

AD _____

Award Number: W81XWH-06-2-0069

TITLE: A Medical Area Network of Virtual Technology (MANVT)

PRINCIPAL INVESTIGATOR: Craig D. Shriver, M.D., FACS, COL, MC

CONTRACTING ORGANIZATION: Windber Professional Services
Windber, PA 15963

REPORT DATE: October 2011

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE 1 Oct 2011		2. REPORT TYPE Final		3. DATES COVERED 22 Sep 2006 - 21 Sep 2011	
4. TITLE AND SUBTITLE A Medical Area Network of Virtual Technology (MANVT)				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-06-2-0069	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Craig D. Shriver, M.D., FACS, COL, MC E-Mail: Craig.Shriver@med.navy.mil				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Windber Professional Services Windber, PA 15963				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
<p>The Clinical Breast Care Project (CBCP) is currently located in three centers, two of which are directly involved in both clinical and molecular research with the other four focused solely on patient accrual. The research centers perform advanced genomic, proteomic and immunology research and tissue banking utilizing state-of-the art technologies. These molecular research technologies yield a vast amount of complex data, conceptually in the terabyte range, particularly as the original data is retained because of the critical nature of the clinical specimens. The complexity, size and importance of clinical data, diagnostic imaging (mammography, ultrasound, PET/CT, MRI and pathology), graphical and digital images that are generated by the high throughput analytical and computational tools already acquired and the need for rapid exchange of clinically relevant information among the centers has required the development of direct links via high speed and high bandwidth networks and the ability to transfer all data generated at the different sites to a central and secure data-warehousing facility. The CBCP centers are fully staffed and operational. Budget for additional sites is under consideration. . The first three years of the MANVT project focused on linking these main facilities via a high speed and dedicated network that enables the secure flow of confidential and de-identified patient data into the project's secure data warehouse. The aim of the MANVT project has been to link, through a dark fiber network, the main CBCP subprojects and satellites to enable rapid and secure clinical and research data uploading, warehousing and querying followed by downloading of the results of successful queries to remote sites with access to the database. The main purpose of this application, to date, has been the development of the Medical Area Network for Virtual Technologies (MANVT), which will be a broadband medical network required to enable the disparate physical sites of the CBCP to be a "virtual linked" single entity. With the successful establishment of this network, the MANVT is evolving to focus on its intended research initiative, to support molecular and clinical research among the sites and evaluate how to use this resource to optimize the delivery of novel, patient centric care. The MANVT will provide the testbed for research into defining this potential for translational medicine, and, in particular, enable its evaluation in both inter-military and civilian-military applications.</p>					
15. SUBJECT TERMS None provided.					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 31	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)

**Final Report
W81XWH-06-2-0069
“Medical Area Virtual Technology”**

Table of Contents

	<u>Page</u>
Cover Page.....	1
SF298.....	2
Table of Contents.....	3
Introductions.....	4
Body.....	4
Key Research Accomplishments.....	7
Reportable Outcomes.....	10
References.....	14
Appendices.....	15

INTRODUCTION

The Clinical Breast Care Project which includes the Windber Research Institute (WRI), the Joyce Murtha Breast Care Center, Walter Reed Army Medical Center and Anne Arundel Medical Center (Maryland) generates large amounts of experimental genomic and cell biology data, images and comprehensive and unique clinical data. The MANVT project was initiated to support the collection, storage, retrieval and sharing of these very large data sets generated by the CBCP at several sites and to assure that the project would have a firm foundation from which to continue into the future. A federated data warehouse has been developed to integrate clinical data, diagnostic imaging and pathology image data as well as data from internal molecular platforms, including immunohistochemistry, fluorescence in situ hybridization, and gene expression, with data from other platforms in the work including copy number variation and DNA sequence. Work is underway to extend the interface of this warehouse to external public domain data for integrative biomedical informatics research. The production version of the data warehouse has been implemented and updated in association with a development partner (InforSense Ltd. (London, UK) now part of IDBS) and a prototype Oracle based image repository system was developed by Concentia Digital (Columbia, MD). These systems are being used internally to support selection of patient cohorts for specific study designs using CBCP data. In addition, these tools have been made available through a secure VPN link to other partners of the collaboration at remote sites. This model has been extended to incorporate data from other congressionally funded programs that are operational at WRI, including the Gynecologic Disease Program (i.e. ovarian, endometrial and cervical cancers) and initial attempts to integrate the Cardiovascular/Diabetes/Obesity programs. The data model has been extended to incorporate these additional data sets using the combined efforts of WRI Biomedical Informatics staff working with InforSense under the existing contract. In addition a fiber optic link between WRAMC and WRI was established and used to transmit large image and other data files for research purposes.

BODY

As discussed below, there are four main areas in which the MANVT project has supported translational research for the CPCB and other translational research initiatives:

1. Connectivity between participating sites;
2. Development of a data warehouse capability for the CBCP and other translational research programs;
3. Implementation of Laboratory Information Management systems LIMs to support various research needs;
- and 4. Development of a robust computational infrastructure for the analysis for large, complex data sets generated by translational research projects that will have the flexibility to grow and evolve with the future needs of the project.

1. Connectivity. The first goal of the MANVT project was to establish a high bandwidth link between to the Walter Reed Army Medical Center and the WRI to facilitate the transfer/exchange of large data files including diagnostic and research images between the two key facilities of the CBCP. Installation of a high speed fiber optic link was

contracted to Fibergate, Inc. and the link was established between WRAMC and WRI. This link was utilized on a regular basis until the CBCP moved from WRAMC to the Walter Reed National Military Medicine Center in Bethesda, MD. in 2011.

2. Data Warehouse capabilities. It is important that the very large amounts of data generated by translational research projects be captured, undergo quality control and are stored/managed in such a way that they can be mined to test and generate new research hypotheses for the project. Critical to the development of such a warehouse is the development of a robust data model driven by the needs of the project.

WRI's data model is unique in that it is a generalized patient-centric, object-oriented data model containing disease independent (such as demographics) and disease-specific objects, including temporal and non-temporal objects. Although initial development and implementation focused on breast disease (as part of the CBCP), the model is applicable to a wide range of diseases under study at WRI including gynecological disease (cancer), cardiovascular disease, diabetes and obesity, because it is non-disease specific representing an abstract level of patient-based modules that serve to integrate and provide clinical and research access to the underlying data streams in a logical manner. Development of the integrated model combining breast disease, from the CBCP (now the Breast Cancer Translational Center of Excellence, BCCoE), and ovarian disease data structures, from the Gynecologic Disease Program (now the Gynecologic Cancer Translational Center of Excellence, GYNCoE) has been completed. This supports the development of research programs looking at the unique linkage between breast and ovarian cancer, which has been noted clinically as a result of possible genetic influences. More than 20 additional questionnaires have been incorporated into the data model to support this effort and reflects input from Georgetown University, University of Pittsburgh Cancer Institute, Walter Reed Army Medical Center and Ohio State University. This further illustrates the flexibility of the underlying data model design and its extensibility towards additional disease/health data sets. Work to implement these extensions to the data model as part of the production data warehouse is currently underway. Commercial organizations, pharmaceutical and biotechnology companies have approached WRI about the potential utilization of the data model and data management system within their enterprises. Additional discussions with TATRC command were undertaken to determine the potential for its use across other TATRC-based projects.

A federated data warehouse has been developed to integrate clinical data, diagnostic imaging and pathology image data as well as data from internal molecular platforms, including immunohistochemistry, fluorescence in situ hybridization, and gene expression, with data from other platforms in the work including copy number variation and DNA sequence. Work is underway to extend the interface of this warehouse to external public domain data for integrative biomedical informatics research. The production version of the data warehouse has been implemented and updated in association with a development partner (InforSense Ltd. (London, UK) now part of IDBS) and a prototype Oracle based image repository system was developed by Concentia Digital (Columbia, MD). These systems are being used internally to support selection of patient cohorts for specific study

designs using CBCP data. In addition, these tools have been made available through a secure VPN link to other partners of the collaboration at remote sites.

Appendix A is a copy of a recently published paper describing our Data Warehouse for Translational Research (DW4TR).

Additional development of the system will directly involve Dr. Jeff Hooke, CBCP, at Walter Reed for use in review and annotation of pathology slides involved in the CBCP program and maintained at WRI's tissue repository. An additional effort was undertaken to develop an electronic tablet for direct entry of breast pathology data to the data warehouse.

3. Laboratory Data Management Systems. A LIMS is important for planning, executing and capturing data from the types of high through-put genomic and proteomic experiments that characterize translational research. The biomedical informatics applications also need to be integrated to the LIMS platform supporting the research informatics. The research informatics and data integration requirements include:

- a powerful and flexible LIMS platform that easily accommodates multiple core technology labs;
- scientific data management applications purpose built for genomics, proteomics and imaging;
- an adaptive platform that is configurable, enabling expansion to multiple science domains;
- an online web collaboration and reporting portal;
- a user-level configuration tool that ensures the solution fits each lab like a custom solution; and
- systematic data capture and contextual data integration.

The research informatics solution needed a robust bioinformatics pipelining system to transform raw data to science results and superior sample traceability from preparation to results, as well as a data marshalling infrastructure to facilitate the export of data to warehouses and analysis packages.

Overall, the informatics vision called for end-to-end connectivity from the distributed collection sites with electronic patient questionnaires, specimen processing, biorepository management and integrated clinical annotations, to query access and specimen requests, lab operations management with automatic data capture, bioinformatics pipelining and integrated clinical and biomolecular results.

To achieve WRI's vision of integrated translational research informatics, the GenoLogics' OMIX life sciences LIMS platform was used as the operational backbone for the specimen handling, laboratory tracking and biomolecular data integration needs. Along with OMIX, the Geneus and Proteus data management applications have been implemented, providing automation and systematic data capture from instruments to

databasing results. LabLink, a web-based collaboration and adaptive reporting portal, was provided to enable easy access to informational views and data.

4. Computational Infrastructure. Another major initiative was a complete upgrade and overall of the IT infrastructure at WRI that supports all of our major programs including the Clinical Breast Care Project, the Integrated Cardiac health Program, and the Gynecologic Cancer Center. The WRI IT infrastructure was nearly 10 years old at the beginning of 2010 and had been built by accretion of hardware without a clear design plan. We began an IT strategic planning exercise by engaging a consulting group, Informatics Management Consultants, LLC, which lead us through this process, helped us develop requirements, organized a vendor conference, aided us in vendor selection and oversaw then installation of the new hardware (see Appendix B and C). We feel that this was a very successful endeavor and the new infrastructure should serve us well over the next 5 years.

KEY RESEARCH ACCOMPLISHMENTS

2010 Accomplishments:

The following are projects had major activity in 2010:

- Electronic Pathology checklist developed by ProLogic. This project moved through the prototype phase and a contract with Proplogic for the completion of this project was put in place (Q1). The application was completed and acceptance plan has been finalized (Q2). Final testing was completed and remaining issues were subitted to ProLogic (Q4).
- Patient/Sample tracking system development with (GenoLogics BMI module) delayed to later in 2010 (Q1). The projects with Genologics were terminated in Q2 and in house development continued through the year.
- CBCP outcome questionnaire/tool development switched from part of the GenoLogics BMI module to development of an In House solution in Q2 and this development continued through the year.
- A gene expression microarray data biological information analysis software was developed by BrainStage, using their proprietary search engine (patent pending) based on the semantic Web (Q4).

Data Warehouse enhancements

- Separating GDP and CBCP data (Role-based security module) Q1, Q2 (completed)
- Sample-centric modules and applications development Q1, Q2 (completed)
- Sample selection projects for different CBCP requests Q1, Q2 (completed)
- Sample-experiment inventory (IHC, gene expression microarray) Q1, Q2 (completed)
- Cleaning modules and data for CBCP/GDP Q1, Q2, Q3, Q4 (completed)
- Developing methods for image archiving in data warehouse Q1, Q2, Q3, Q4 (ongoing)
- Re-modeling of revised GCC questionnaires Q1, Q2, Q3 (completed)

- Implementing new GCC questionnaires Q2, Q3 (completed)
- Implementing new GCC questionnaires tested and performed QA using artificially generated data. QA with the PI/clinical domain experts pending. Q4

Portal development User interface to data warehouse (based on InforSense VisualSense 5.0.1)

- 1) Design and develop portal for biomedical needs (general structure, security, management, visualization) – complete
- 2) Move all files/links/tools to the portal for public use – on going
- 3) Publish workflows to the portal so scientists could analyze data without modifying workflows - complete
- 4) Link between sample and pathology check list as well as between sample and core questionnaire (will be one of first projects published to the portal) – complete
- 5) Patient cohort selection using portal filters (plan to create workflow that step by step will allow selection of the cohort using GUI interface) – complete

Data Analysis

- 1) IHC biomarker analysis for CBCP
- 2) Microarray analysis (blood) for CBCP
- 3) Racial disparity (clinical and pathology) data for CBCP.

Strategic IT infrastructure plan for the CBCP—all are already completed.

- Q1 - A consulting group, Informatics Management Consultants, LLC, has been engaged to develop a strategic IT infrastructure plan for the CBCP—complete
- Q2 - IT infrastructure analysis – complete
- Q2 - Design new IT infrastructure which contains: virtual servers; virtual storage; compute farm; and network hardware – complete
- Q2 - Develop a RFP (Request For Proposal) to have new IT infrastructure implemented - complete
- Q2 - Distribute a RFP to ten local vendors - complete
- Q3 - Hold vendor conference – complete
- Q3 - Vendor conference was held
- Q3 - Vendor has been selected (Link Computer Corporation)
- Q3 - Developed final plan with Vendor
- Q3 - Beginning procurement
- Q4 -Procurement of all hardware complete
- Q4 - Hardware installation complete, which includes network hardware, servers, and virtual storage
- Q4 - Network hardware configuration complete
- Q4 - Cut-off of WRI from the MANVT network complete and now Level 3 providing internet connectivity for WRI
- Q4 - Virtualization of current physical servers – complete

2011 Accomplishments:

In 2011 we put a special focus on re-structuring the IT architecture for CBCP and WRI operations. Since its inception the IT infrastructure was growing with the expansion of CBCP/WRI activities. We have reached a point of time that the whole system needs to be completely re-structured to lay the foundation for future research and support activities of the next phase of CBCP and GCC when they become part of the Military Cancer Center of Excellence. The following are the main activities:

- Procurement of all hardware complete
- Hardware installation complete, which includes network hardware, servers, and virtual storage
- Network hardware configuration complete
- Cut-off of WRI from the MANVT network complete and now Level 3 providing internet connectivity for WRI
- Virtualization of current physical servers – on going

Other 2011 activities are listed below:

Closing of two contracted specific projects:

- Electronic Pathology checklist developed by ProLogic. Final testing is complete and remaining issues are submitted to ProLogic.
- A gene expression microarray data biological information analysis software was developed by BrainStage, using their proprietary search engine (patent pending) based on the semantic Web

Data Warehouse enhancements:

- Cleaning modules and data for CBCP/GCC – completed.
- Re-modeling of revised GCC questionnaires – completed.
- Implementing new GCC questionnaires – on going and close to be finished with QA using artificially generated data. QA with the PI/clinical domain experts pending.

Portal development User interface to data warehouse (based on InforSense VisualSense 5.0.1):

- Move all files/links/tools to the portal for public use – complete
- SNP analysis using GeneSense – terminated
- Microarray analysis using InforSense workflow builder – terminated
- The Data Warehouse paper published
- The paper on using the CBCP data warehouse resubmitted

Data Analysis with several abstracts published and manuscripts in preparation:

- IHC biomarker analysis for CBCP – on going
- Microarray analysis (blood) for CBCP – on going
- Racial disparity (clinical and pathology) data for CBCP – on going
- Clinical tissue blocks inventory for CBCP – on going

REPORTABLE OUTCOMES

Papers and manuscripts:

Maskery, S., Zhang, Y., Jordan, R., Hu, H., Hooke, J., Shriver, C., & Liebman, M. 2006e, "Co-Occurrence Analysis for Discovery of Novel Breast Pathology Patterns", *IEEE Transactions on Information Technology in Biomedicine 2006*, vol. 10, no. 3, pp. 497-503.

Xiang, G., Liu, R., Shriver, C., Hu, H., & Liebman, M. 2006, "Assessing Semantic Similarity Measures for the Characterization of Human Regulatory Pathways", *Bioinformatics Oxford Journals*, vol. Bioinformatics, doi:10.1093/bioinformatics/bt1042.

Yang Y, Heckman, C., Shriver, C., Becker, T., Liebman, M., & Brzeski, H. 2006, "Gene Expression Profiles in White Blood Cells of Patients with Breast Cancer - Correlation with Detectability of Tumor by Mammographic Screening", *Breast Cancer Res Treat*, vol. 94, no. no. Suppl 1, p. S59-S60.

Yang, S., Guo X, Yang Y, Papcunik, D., Heckman, C., Hooke, J., Shriver, C., Liebman, M., & Hu, H. 2006, "Detecting Outlier Microarray Arrays by Correlation and Percentage of Outliers Spots", *Cancer Informatics* no. 2, pp. 351-360.

Deng, H., Liu, Y., Shen, H., & Hu, H. 2007, "An Introduction to Biomedical Informatics. Current Topics in Human Genetics: Studies of Complex Disease. [in press]," H. Deng et al., eds., World Scientific Publishing.

Hu, H., Mural, R., Liebman, M., & (Editors) 2008, *Biomedical Informatics in Translations Research* Artech Publishing House.

Maskery, S., Hu, H., Hooke, J., Shriver, C., & Liebman, M. 2008, "A Bayesian Derived Network of Breast Pathology Co-occurrence", *Journal of Biomedical Informatics* no. 41, pp. 242-250.

Kvecher L, Wu W, Hooke JA, Shriver CD, Mural RJ, and Hu H. An Approach to Correlate the Temporal Information To Facilitate Specimen Selection in the Breast Cancer Research Project. *Cancer Research* 2009;69;763s.

Zhang Y, Kvecher L, Sun W, Gutchell EM, Mural RJ, Shriver CD, Liebman MN, and Hu H. A quality assurance issue tracking tool to facilitate the enhancement of clinical data quality (submitted).

Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, Schwab G, Gao D, Gao J, Kubatin V, Shriver CD, Hooke JA, Maxwell LG, Kovatich AJ, Sheldon JG, Liebman MN, and Mural RJ. DW4TR: a data warehouse for translational research. *Journal of Biomedical Informatics* 2011;1004-1019.

Invited/Conference presentations:

Maskery, S., Hu, H., Hooke, J., Shriver, C., & Liebman, M. "Validation of Bayesian Network Analysis of Breast Disease Co-Occurrence - Implication for Future Application to Clinical, Molecular, and Epidemiological Data Generated by the Clinical Breast Care Project [Podium Presentation]", Podium Presentation at the 7th Annual CBCP Offsite Meeting, November 2006, Cumberland, MD.

Hu, H. "A Data Warehouse for Translational Research (Podium Presentation)", Podium Presentation at the Translational Health: The Next Generation of Medicine Conference, Pasquerilla Conference Center, Johnstown, PA. 2006

Hu, H., Correll, M., Sheldon, J., & Liebman, M. "Development of an Oracle-based Integrative Biomedical Informatics Data Warehouse (Podium Presentation)", Podium Presentation at the Oracle Life Science Users Group Meeting, April 3, 2006, Boston, MA.

Liebman, M. & Mural, R. "Translational Medicine: Defining it's Role and Future (Podium Presentation)", Podium Presentation at the Translational Health: The Next Generation of Medicine Conference, Pasquerilla Conference Center, Johnstown, PA. 2006

Hu, H. "Unlock Insight from Your Clinical Data [Podium Presentation]", Joint World-Wide WebEx by InforSense and the Windber Research Institute, September 19, 2007.

Hu, H. "Application of Biomedical Informatics to Translational Research [Invited Speaker]", Invited Speaker at The First Annual EITC-Bio Workshop, Princeton University, Princeton, NJ.

Hu, H. "Synergy of Informatics and Biomedical Research in Academia and Industry [Invited Panelist]", Invited Panelist at the First Annual EITC-Bio Workshop, June 7, 2008, Princeton University, Princeton, NJ.

Maskery, S., Zhang, Y., Hu, H., Shriver, C., Hooke, J., & Liebman, M. "Caffeine Intake, Race, and Risk of Invasive Breast Cancer, Lessons Learned from Data Mining a Clinical Database (Podium Presentation and Paper)", Podium Presentation and Paper, Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, UT, June 22-23, 2006, pp. 714-718.

Mural, Richard J. Panel Member, "Informatics Tools to Enable Integrative Translational Research", 2009 AMIA Summit on Translational Bioinformatics, San Francisco, CA, March 15-17, 2009.

Mural, Richard J., DeGreef, James. "Accelerating Translational Research with a Comprehensive Informatics Solution that Connects Discovery and Clinical Data", 2009 AMIA Summit on Translational Bioinformatics, San Francisco, CA, March 15-17, 2009.

Hu H. October 13, 2009. Biomedical Informatics in Translational Research. InnovationWell Annual Meeting, Bryn Mawr College, Philadelphia, PA.

Hu H. October 22, 2010. *The development of the Data Warehouse for Translational Research (DW4TR)*. Invited lecture hosted by: Dr. Mike Sierk, Saint Vincent University, Latrobe, PA

Hu H. April 12-14, 2011. *A Data Warehouse for Translational Research*. The 10th Annual BIO-IT World Conference and Expo. Boston, MA.

Conference posters presentations:

Maskery, S., Hu, H., Mural, R., Hooke, J., Shriver, C., & Liebman, M. "Separate DCIS Progression Pathways Derived from Breast Disease Heterogeneity Data (Poster Presentation)", Poster Presentation at the American Society of Clinical Oncology Annual Meeting, Atlanta, GA, June 2-6, 2006.

Maskery, S., Zhang, Hu, H., Hooke, J., Shriver, & Liebman, M. "Breast Pathology Co-Occurrence in Stratified Populations, Implications for Breast Cancer Development in Different Populations [Poster Presentation]", Poster Presentation at the 2006 American Association for Cancer Research Annual Meeting, Washington, DC. April 1 - 5, 2006.

Bekash, A., Maskery, S., Kvecher, L., Hooke, J., Liebman, M., Shriver, C., Mural, R., & Hu, H. "A Pilot Study of Controversial Breast Cancer Risk Factors Using the Clinical Breast Care Project Database as a Research Environment [Poster Presentation]", Poster Presentation at the 30th Annual San Antonio Breast Cancer Symposium, December 13-16, 2007, San Antonio, TX.

Maskery, S., Hu, H., Liebman, M., Hooke, J., Shriver, C., & Taioli E "Traditional Breast Cancer Risk Factors and Common Breast Pathologies in Post-Menopausal Women [Poster Presentation]", Poster Presentation at the 30th Annual San Antonio Breast Cancer Symposium, December 13-16, 2007, San Antonio, TX.

Maskery, S., Hu, H., Liebman, M., Shriver, C., Verbanac, K., Tafra, L., & Rosman, M. "Bayesian Analysis of Recurrence in Lymph Node Positive and Lymph Node Negative Breast Cancer Patients", Poster Presentation at the 30th Annual San Antonio Breast Cancer Symposium, December 13-16, 2007, San Antonio, TX.

Yang, S., Guo, X., & Hu, H. 2007, "An R Function to Detect Microarray Outlier Slides", *Genomics, Prot and Bioinformatics* p. Revision Submitted.

Zhang, Y., Correll, M., & Hu, H. 8 A.D., "Chapter 9. Data Analysis," in *Biomedical Informatics in Translations Research*, H. Hu, R. Mural, & M. Liebman, eds., Artech Publishing House.

Zhang, Y., Sun, W., Gutchell, E., Mural, R., Shriver, C., Liebman, M., & Hu, H. "An Issue Tracking System to Facilitate the Enhancement of Clinical Data Quality in the Clinical Breast Cancer Project [Poster Presentation]", AMIA 2007 Annual Symposium November 10-14, 2007, Chicago, IL.

Kvecher L, Wu W, Kohr J, Shriver CD, Mural RJ, and Hu H. DCU: A Data Correction Utility to Correct Clinicopathologic Data in a Data Warehouse. AMIA 2011 Annual Symposium, October 22-26, 2011, Washington, DC.

REFERENCES

N/A

APPENDICES



DW4TR: A Data Warehouse for Translational Research

Hai Hu^{a,*}, Mick Correll^{c,1}, Leonid Kvecher^a, Michelle Osmond^{b,2}, Jim Clark^{b,2}, Anthony Bekhash^a, Gwendolyn Schwab^a, De Gao^{d,2}, Jun Gao^{d,2}, Vladimir Kubatin^{c,2}, Craig D. Shriver^e, Jeffrey A. Hooke^e, Larry G. Maxwell^e, Albert J. Kovatich^f, Jonathan G. Sheldon^{b,3}, Michael N. Liebman^{a,4}, Richard J. Mural^a

^a Windber Research Institute, Windber, PA, USA

^b InforSense LLC., London, UK

^c InforSense Ltd., Boston, MA, USA

^d InforSense Ltd., Shanghai, China

^e Walter Reed Army Medical Center, Washington, DC, USA

^f MDR Global, Windber, PA, USA

ARTICLE INFO

Article history:

Received 29 September 2010

Accepted 4 August 2011

Available online 22 August 2011

Keywords:

Data Warehouse

Translational research

Patient-centric data model

User interface

Ontology

ABSTRACT

The linkage between the clinical and laboratory research domains is a key issue in translational research. Integration of clinicopathologic data alone is a major task given the number of data elements involved. For a translational research environment, it is critical to make these data usable at the point-of-need. Individual systems have been developed to meet the needs of particular projects though the need for a generalizable system has been recognized. Increased use of Electronic Medical Record data in translational research will demand generalizing the system for integrating clinical data to support the study of a broad range of human diseases. To ultimately satisfy these needs, we have developed a system to support multiple translational research projects. This system, the Data Warehouse for Translational Research (DW4TR), is based on a light-weight, patient-centric modularly-structured clinical data model and a specimen-centric molecular data model. The temporal relationships of the data are also part of the model. The data are accessed through an interface composed of an Aggregated Biomedical-Information Browser (ABB) and an Individual Subject Information Viewer (ISIV) which target general users. The system was developed to support a breast cancer translational research program and has been extended to support a gynecological disease program. Further extensions of the DW4TR are underway. We believe that the DW4TR will play an important role in translational research across multiple disease types.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Translational research requires integration of clinicopathologic data, biospecimen data, and molecular data, across multiple data collection and generation platforms [1–4]. It is also critical to make these data usable at the point-of-need. From the data management perspective, clinicopathologic data and molecular data have distinct features. The former are characterized by a large number of data fields, often with missing values due to failure to collect, non-existence or inaccessibility of the data. Molecular data, on the other hand, are characterized by a limited number of data

fields, large number of records, and ever-evolving data types associated with the development of new technologies. While a translational research project typically collects and generates its own clinicopathologic and molecular data, public domain data are often used as well. Furthermore, temporal information needs to be properly managed and presented.

Many clinical and translational research projects use questionnaires as a clinical data collection instrument, and currently using data from Electronic Medical Records (EMRs) for translational research is gaining momentum as EMR systems replace paper-based medical records. Compared to EMRs, questionnaires are simpler as only questions useful to the specific research program of the disease would be included, thus managing such data does not require a comprehensive system covering all the human diseases. An EMR system cannot be used as a data management system for translational research, due to the transactional nature of the former and the reporting and analysis requirements of the latter. The requirements from the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) for protection of patients' Protected Health

* Corresponding author. Address: Windber Research Institute, 620 7th St., Windber, PA 15963, USA. Fax: +1 814 467 6334.

E-mail address: h.hu@wriwindber.org (H. Hu).

¹ Present address: Dana-Farber Cancer Institute, Boston, MA, USA.

² Present address: IDBS, Burlington, MA, USA.

³ Present address: Oracle, Swindon, UK.

⁴ Present address: Strategic Medicine, Kennett Square, PA, USA.

Abbreviations

2D-DIGE	2-Dimensional, Difference In Gel Electrophoresis	IT	Information Technology
ABB	Aggregated Biomedical-Information Browser	LC-MS	Liquid-Chromatography Mass-Spectrometry
aCGH	Array Comparative Genomic Hybridization	LIMS	Laboratory Information Management System
AJAX	Asynchronous JavaScript and XML	MeSH	Medical Subject Heading
AJCC	American Joint Committee on Cancer	MIAME	Minimum Information About a Microarray Experiment
ASCO	American Society of Clinical Oncology	MS	Mass Spectrometry
BioPAX	Biological Pathway Exchange	NBIA	National Biomedical Imaging Archive
BMI	Body-Mass Index	NCI	National Cancer Institute
caBIG [®]	Cancer Biomedical Informatics Grid	NCICB	National Cancer Institute Center for Bioinformatics
caBIO	Cancer Bioinformatics Infrastructure Objects	NIH	National Institutes of Health
caCORE	cancer Common Ontological Reference Environment	OCT	Optical Cutting Temperature medium
caDSR	cancer Data Standards Registry and Repository	OLAP	On-Line Analytical Processing
CAP	College of American Pathologists	MOLAP	Multi-dimensional OLAP
CBCP	Clinical Breast Care Project	ROLAP	Relational OLAP
CDW	Clinical Data Warehouse	OWL	Web Ontology Language
CGEMS	Cancer Genetic Markers of Susceptibility	PHI	Protected Health Information
CTSA	Clinical and Translational Science Award	PI	Principal Investigator
DICOM	Digital Imaging Communications in Medicine	PR	Progesterone Receptor
DNA	Deoxyribonucleic Acid	QA	Quality Assurance
DW	Data Warehouse	RAID	Redundant Array of Independent Disks
DW4TR	Data Warehouse for Translational Research	RAM	Random-Access Memory
EAV	Entity-Attribute-Value	REMBRANDT	Repository of Molecular Brain Neoplasia Data
EMR	Electronic Medical Record	RDBMS	Relational Database Management System
ER	Estrogen Receptor	RIN	RNA Integrity Number
ETL	Extract, Transform, and Load	RNA	Ribonucleic Acid
FF	Flash Frozen	RT-PCR	Real-Time Polymerase Chain Reaction
FFPE	Formalin-Fixed, Paraffin-Embedded	SAS	Statistical Analysis System
FISH	Fluorescence In Situ Hybridization	SNOMED-CT	Systemized Nomenclature of Medicine-Clinical Terms
GB	Giga-Byte	SNP	Single-Nucleotide Polymorphism
GDP	Gynecological Disease Program	SPSS	Statistical Package for the Social Sciences
GO	Gene Ontology	STRIDE	Stanford Translational Research Integrated Database Environment
H&E	Hematoxylin and Eosin	TB	Tera-Byte
HER2	Human Epidermal growth factor Receptor 2	UML	Unified Modeling Language
HIPAA	Health Insurance Portability and Accountability Act of 1996	UMLS	Unified Medical Language System
HL7	Health Level 7	VCDE	Vocabulary and Common Data Elements
I2B2	Informatics for Integrating Biology and the Bedside	VPD	Virtual Private Database
ID	Identification	VPN	Virtual Private Network
IHC	Immunohistochemistry	WRAMC	Walter Reed Army Medical Center
IRB	Institutional Review Board	XML	Extensible Markup Language
ISIV	Individual Subject Information Viewer		

Information (PHI) also differ between the two systems. In addition, translational research may not need all of the data contained in an EMR and it requires data from other sources as well including molecular data. It is important that a data centralization system be developed for translational research capable of inputting clinical data from questionnaires and EMRs for any human disease, and integrating molecular data from different biochemical, genomic and proteomic experimental platforms.

Data warehousing as an important component of biomedical informatics, provides a way for data integration in clinical and translational research [2,5–13]. A Data Warehouse (DW) is an information repository for data analysis and reporting [14,15]. For raw clinical data, an Entity-Attribute-Value (EAV) data model is often used due to their sparseness and dynamic nature [16–18]. Such data models require the development of ontologies to manage the complex relationships between clinical concepts/questions and possible answers. The user interface is often provided by an On-Line Analytical Processing (OLAP) tool, which provides multidimensional data analysis capabilities by aggregating data across hierarchically organized dimensions using either a multidimensional (MOLAP) or relational (ROLAP) model [14,15,19]; the former

relies on pre-computed data cubes whereas the latter directly queries a relational database.

Developing a comprehensive DW system to meet the needs of translational research faces many challenges. Based on our own experience, such challenges include; management of operationally important information, such as HIPAA compliance and PHI, data input from questionnaires with version control, data input from EMRs, biospecimen collection and banking, data de-identification, molecular images, multiple-platforms of molecular data, temporal information, and data ownership, etc. From the system development perspective challenges include; development and application of ontologies including controlled vocabulary and modeling of clinical data, molecular data, image data, and temporal data. In addition, development of graphical user interfaces for different levels of users is important, and data security and accessibility should be properly addressed.

Ontology development is very important to data integration and exchange; for the purpose of this paper we focus on its physical artifacts of terminologies/controlled vocabulary and information model/data model [20]. Many ontologies have been developed, each serving a specific purpose. In the clinical field,

commonly accepted standards/ontologies include; Health Level 7 (HL7) for exchange of clinical messages to support clinical practice and healthcare service [21]; Medical Subject Heading (MeSH) for medical literature indexing which has also proven useful for classification of biomedical entities [22]; Systemized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), as a comprehensive clinical healthcare terminology organized hierarchically [23]; and Digital Imaging Communications in Medicine (DICOM) for medical image annotations [24]. For biological research, Gene Ontology (GO), arguably the most successful biological ontology, is aimed to standardize the representation of gene and gene product attributes across species and databases [25]; Minimum Information About a Microarray Experiment (MIAME) for microarray-based technologies is generally accepted by the community [26]; Biological Pathway Exchange (BioPAX) is for pathway data exchange [27]. NCI Thesaurus provides reference terminology for many NCI and other systems covering vocabulary for clinical care and translational and basic research [28]. Systems have been developed to unify existing ontologies or provide ontology library services, for example the BioPortal by the US National Center for Biomedical Ontology [29], and the Ontology Lookup Service by European Bioinformatics Institute based on the Open Biomedical Ontology format [30]. The Unified Medical Language System (UMLS) developed by the US National Library of Medicine, aims to provide ad hoc linkages across life science ontologies [31]. A number of reviews are available including ones discussing development, implementation, and use of ontologies [20,32–34].

Temporal information is also very important in modeling health and disease development. We consider disease as a process, not a state. Breast cancer, for example, takes many years to develop before a lesion can be detected by mammography [35]. First considered by artificial intelligence researchers in the 1980s, temporal reasoning and temporal data query became a topic of interest in biomedical/medical informatics in the 1990s [36–38]. To enable “controlled language use” of the temporal information in healthcare, a European Prestandard CEN ENV 12381 has been developed [39]. Temporal information needs to be collected, modeled, and presented. Collection of temporal information can be done explicitly through questionnaires or data forms which record historical information, or abstracted through the time-stamps of data collected and analyzed over time. For the latter, temporal-abstraction and temporal-querying systems have been developed which have proven clinically useful in monitoring clinical disorders [40–42]. A data model comprising Parameters, Events, and Constants has been defined for a temporal query language [43,44], where Parameters are basically entities with values bearing a time-stamp, Events are external acts act upon a patient which could be time-stamped or occur over a time interval, and Constants are non-temporal measures. There have been reports of a data management or DW system incorporating temporal information for clinical trials, time-series gene expression microarray experiment, and medical information events in a hospital setting [38,45,46].

The many challenges faced in developing a comprehensive DW system for translational research make this a difficult undertaking. In practice, different systems have been developed focusing on different challenges based on specific needs of supported program(s). For example, I2B2, Informatics for Integrating Biology and the Bedside [47], took a top-down “Enterprise” approach with an open architecture enabling integration of outside modules performing specialized functions that are needed for translational research [48,49]. It address a very important and specialized niche in the clinical informatics space, namely, providing researchers with sufficient access to clinical data to enable study design and cohort selection, while minimizing many of the patient privacy risks and concerns [50–54]. A central conundrum of translational research is that access to detailed clinical data typically requires proper

informed consent and IRB approval; however, in order to design a cohort study or to determine if a particular project is even feasible, a researcher often needs access to aggregate statistics for key clinical attributes of interest at an early stage of the process. I2B2 overcomes these challenges by enabling researchers to design and execute queries against a de-identified data mart that return only aggregate counts; furthermore, results are subtly obfuscated to avoid identification by more sophisticated combinations of queries. Having been deployed for a number of projects across multiple sites, I2B2 is a mature software offering. However, it does not currently provide the level of detailed analysis necessary for clinical and translational research beyond cohort selection and study design [51–54]. STRIDE, The Stanford Translational Research Integrated Database Environment [55], focuses on a Clinical Data Warehouse (CDW) and supports a virtual biospecimen model for accessing several specialized tissue banks [13,56,57]. It also has a Research Database management system supporting multiple logically separate research databases. An EAV model is used for data storage and HL7 Reference Information Model is used for data representation. Its semantic layer consists of a framework supporting multiple terminologies. Some of the Clinical and Translational Science Award (CTSA) centers [61] also have data warehousing efforts.

There are also data warehousing efforts in the cancer Biomedical Informatics Grids (caBIG®) [62]. caBIG® is an information network enabling members of the cancer community to share data and knowledge based on a core Grid architecture. It was launched in 2003 by the NCI Center for Bioinformatics (NCICB) of the US NIH [63,64]. One of its data warehousing efforts is caBIO, a major component of caCORE (cancer Common Ontological Reference Environment) that was developed before the launch of caBIG® [65], enabling access of biomedical annotations from curated data sources in an integrated view. As one of the released products of caBIG®, the current version caCORE 3.x includes Enterprise Vocabulary Services for hosting and managing vocabulary, cancer Data Standards Registry and Repository (caDSR) for hosting and managing metadata, and the caCORE Software Development Kit [66]. Another data warehousing approach for caGrid is a semantic web-based data warehouse for creating relationships among caGrid models, accomplished through the transformation of semantically-annotated caBIG® Unified Modeling Language (UML) information models into Web Ontology Language (OWL) ontologies that preserve those semantics [67]. Still another caBIG effort, caIntegrator [58], is a framework for data integration currently capable of integrating microarray data in caArray and imaging data from the National Biomedical Imaging Archive (NBIA). It has been applied to several projects, e.g., REMBRANDT [59] and CGEMS [60]. These efforts are best described as development projects using the caIntegrator framework. None of these efforts in caBIG, however, enables users to dynamically interrogate the clinicopathologic data in a multidimensional manner by non-informatics specialists using an intuitive interface.

We began our DW development focused on a questionnaire-based system, but later realized that a patient-centric and expandable data model would be a better solution. We envisioned developing a system to support translational research in general and began by developing a system to support the Clinical Breast Care Project (CBCP) [68] but designed it to be expandable to support additional programs. This approach has been validated by our successful expansion of the system to support a second translational research program, the Gynecological Disease Program (GDP) [69]. It is currently under further expansion to support a third translational research program, a consortium effort headed by Thomas Jefferson University involving five organizations which is characterizing a set of 5000 invasive breast cancer cases. All these projects have multiple platforms for clinical and molecular studies.

Development of our system, the Data Warehouse for Translational Research (DW4TR), focused on three challenges in DW development. The first was the development of a pragmatic data model to satisfy the needs of questionnaire-based clinical data input. Despite the existence of a large number of clinical and biological ontologies, there are often gaps, in supporting a research project, between the available ontologies and the specific project needs [70,71]. Many groups therefore create their own ontologies [72,73]. Given the characteristics of clinical and molecular data, it is more challenging to develop a general and flexible system for clinical data; therefore we have made this our first priority [9]. In our case, both CBCP and GDP utilizes not only data elements covered by existing ontologies, but also detailed and specific data elements that are currently not covered. Comparing what we need to what is available, we concluded that completely adopting one or two existing ontologies for our DW development is not feasible (c.f. Section 5). Therefore, we developed our own framework ontology but referred to existing ones, to satisfy our immediate needs first, planning to map our ontology to existing ontologies at a later time to enable transportability.

The second challenge, we focused on was interface development. A user-friendly interface tailored to the non-informatics specialist is needed to ensure both the utility and adoption of this system. The interface that we developed is composed of an Aggregated Biomedical-Information Browser (ABB) and an Individual Subject Information Viewer (ISIV). ABB is ROLAP-based, thus all the calculations and queries are performed at the time of use rather than relying on a pre-calculated data cube, enabling clinicians and scientists who are not informatics experts to dynamically interrogate any combination of the myriad of data elements stored in the DW4TR. The third challenge we focused on was the management of temporal information. We defined three temporal data types and applied them to data attributes to enable the use of the temporal information in the DW4TR from both interfaces. In addition to focusing on these three challenges, we also addressed a number of other challenges described earlier. For example, in managing de-identified clinical information of CBCP and GDP, we developed solutions to satisfy different requirements for data security and accessibility that is also described.

2. Methods

2.1. System design and development

Initially developed to support the immediate needs of managing the clinical, biospecimen, and molecular data for the CBCP, our design of the DW4TR to be flexible and extensible enabled subsequent expansion to support GDP. Both programs use questionnaires for collecting clinicopathologic information, making expansion of the light-weight production system straight forward.

2.1.1. The data model

The EAV model is used for the original raw clinical data which are subsequently hosted in the DW in an extensible data model reflecting the ontology, where the relationships among the attributes are presented in a meaningful way to the users through a hierarchical organization. The extensible data model is currently composed of the following components:

2.1.1.1. Patient-centric clinical data model. The clinical data model was developed to reflect the physician–patient interaction. The ontology framework was developed by a multi-disciplinary team composed of research physicians, biomedical informaticians, IT developers, surgeons, pathologists, and gynecologists and was based on the standard clinical workflows and clinical experiences.

The development not only referred to existing standards and ontologies mainly MeSH and SNOMED-CT but also to the NCI Thesaurus and caBIG VCDE [22,23,28,74], reflecting other reported methods and efforts [32,33,75]. This ontology is reflected on the physical data model, which is hierarchical and composed of a number of primary modules, e.g., “Diagnostic”, which are made up of secondary modules (e.g., “Biopsy”) and attributes. A secondary module is further composed of tertiary modules and attributes, and so on. An attribute is a fine-grained object that is composed of the name, the data type, and the temporal characteristics (see below). A simple attribute is often called a data element, e.g., sex. A complex attribute is composed of multiple facets, e.g., exercise is composed of frequency, intensity, duration, and type. This way, each data element of a study is represented in the model either by an attribute or by a sub-module. The sub-modules are structured hierarchically to reflect the ontology.

2.1.1.2. Specimen-centric molecular study data model. Molecular (biochemical, genomic and proteomic) analysis are performed on collected specimens using multiple experimental platforms. Every experiment is done using biospecimens, including body fluids, solid tissues, and their derivatives. Thus, a specimen-centric data model is a natural choice for these types of data, which are connected to the clinical data model hierarchy as a series of sub-modules. We initially focused on immunohistochemistry (IHC), Fluorescence In Situ Hybridization (FISH), and gene expression microarray data and adopted existing clinical or molecular data standards and guidelines [26,76,77]. The storage and retrieval of high throughput molecular data poses a number of IT challenges. A result file from a single microarray can be tens of megabytes, and it is not uncommon for a single experiment to make use of hundreds of such arrays. Common approaches for handling this type of data include: organizing output files into a hierarchical file system, loading of raw or partially transformed data into a DW, or a hybrid approach of maintaining certain metadata in a database but referring back to original output files with file pointers. Given the large number of molecular modalities used in our current and future research, no single method was considered suitable, therefore our system was designed to take an “all of the above approach” that makes use of flexible data processing pipelines. With this approach distinct pipelines can be created across or even within assay types and all of the storage methods discussed above can be supported. This approach has enabled us to optimize storage of each assay group individually, as well as provide a mechanism for handling changing technology and legacy data.

2.1.1.3. Temporal data model. Critical to human disease studies is the proper representation of temporal information. We define three types of temporal data; (1) Static, which is data with no temporal dimension, e.g., ethnicity. (2) Event, which is associated with a specific time point, e.g., a surgical procedure. (3) Interval, which is associated with a starting and ending time point, e.g., a course of medication. Each attribute in the data model is tagged with a proper temporal status and populated with proper values for temporal information representation and analysis.

2.1.1.4. Medical image data model. Medical imaging is an important part of clinical practice and a source of data for various studies. We initially focus on digitized and digital mammograms, which follows the DICOM standards [24] and we de-identified the images using DICOM Anonymizer Pro. The data files associated with these images are large and they impose a challenge in data management, e.g., storing those image files in the database of the DW will drastically impede the system performance. Besides, analyzing images typically requires specialized software. Thus, in the DW4TR, we store the image files on a file server; the file locations are then

stored in the DW and logically linked to the patient. The annotations and a thumb nail of the image are also stored in the database to enable efficient search of relevant features. We intend to apply the same principle to other images.

2.1.1.5. Relationships between questionnaires and the data model. The data for both CBCP and GDP studies are collected using questionnaires each having multiple questions. Each question has one or more data elements. The data elements could contain values and temporal information. If a data element contains a value, it is mapped to an attribute(s). Some questions ask about the same event at different time points, and these questions are mapped to the same attribute with corresponding temporal information. There are also different questionnaires collecting the same information, e.g. for different studies, and they have been mapped to the same attributes as well. Finally each attribute is mapped to the data model as described above.

2.1.2. User interface development

The requirement to develop an intuitive, user friendly system that could be used directly by clinicians and researchers as opposed to being a specialist tool useful only to the technically savvy, led to our decision to take the dimensional modeling approach. While traditional query interfaces are considered intuitive by IT professionals, we found that the dimensional modeling approach was generally preferred by our target demographic. A number of criteria were considered in designing the physical data structure to support the clinical data warehouse, the first of which was the ability to handle a large number of dimensions. A custom binning capability that is convenient to use, was an important feature requested by the users. As is the case with any longitudinal study, the temporal dimension of data is critical, and therefore proper handling of the temporal dimension was another important concern. Efficient handling of sparsely distributed data and support for multiple terminology sets were also important criteria. During such comprehensive communications with the end users, we concluded that all the desired analyses could be categorized as either an aggregate data analysis or an individualized data analysis. Thus, we designed a ROLAP-based ABB for the former, and an ISIV for the latter.

2.2. Implementation

2.2.1. Development environment

Two DELL Windows server systems, one for the application and the other for the database, were used. Both servers have two Quad Core Intel® Xeon® processors and 8 GB RAM. The database server hard drive is 1 TB and for the application server 0.5 TB, both are in the RAID 5 configuration. Development is based on the Oracle RDBMS 10 g, using the InforSense platform that enables a series of analyses to be performed in an analytical workflow environment with each analysis step represented as a node in the workflow [78]. The user-interface is Java based and web accessible. The production system has the same configuration. The DW4TR also supports a one-server system configuration.

2.2.2. Data model and interface implementation

Our physical data model is composed of two distinct elements, an attribute repository, and an attribute ontology. The attribute repository stores data in the EAV model. The attribute ontology is a separate hierarchical data model capturing and providing structure to the relationships among the attributes that have been collected so that the ontology (logical patient-centric or specimen-centric data model) are physically positioned in a hierarchical organization of attributes for presentation in a meaningful way to the users of the system.

In implementing the data model, first the source questionnaires are reviewed and a logical mapping of the data attributes onto the newly designed patient modules is developed. This process requires the definition of the data variable type and other metadata for different data attributes, and often requires splitting or merging data elements in order to accommodate the difference in requirements of capturing the data versus analyzing the data.

Parallel to the development of the physical and logical data model was the development of the end user interfaces. Our system provides users with two distinct applications: the ABB, and the ISIV. These Java tools are entirely web based with zero client side foot print, and do not require any additional browser plug-ins. They provide users with a highly interactive and dynamic experience, making use of Asynchronous JavaScript and XML (AJAX) techniques for asynchronous rendering and updating. Both tools are deployed within the InforSense web portal environment, which provides the services used for data source connection pooling, user authentication and authorization, as well as web session management.

2.2.3. Data loading

Multiple sources of data in different format are loaded into the DW4TR, including direct Oracle-to-Oracle load from the Laboratory Information Management System (LIMS) and flat files [12,79]. Data loading is performed through a standard process referred to as Extract, Transform, and Load (ETL) [14,15].

2.3. Data sources

The current DW4TR is host to clinical data from CBCP and GDP. Subjects are enrolled into the programs via HIPAA compliant, Institutional Review Board (IRB) approved protocols at multiple participating clinical sites, and these protocols explicitly specified WRI as the data integration center for the programs. Clinical data are collected through multiple questionnaires, and biological specimens are collected, processed, banked, and analyzed using genomic and proteomic experimental technologies.

CBCP and GDP happen to be two distinct types of programs in administrative structure, clinical dataset, and tissue banking. They have different data security and access requirements, commanding distinct solutions in the DW4TR. CBCP is a centralized program using one set of the questionnaires, one LIMS, one Tissue Bank, and one data warehouse (the DW4TR). All the participating organizations use approved master IRB protocols from the Walter Reed Army Medical Center (WRAMC), with minor adaptations to the requirements of the local institution for local IRB approval, which is subsequently approved by the IRB of the US Army Medical Research and Materiel Command that contracted the Henry Jackson Foundation to manage the CBCP. Proper paperwork was done to enable centralized tissue banking and data warehousing at WRI. The IRB approved questionnaires contain certain date information such as date of birth which patients are consented to, thus the CBCP clinical data is a "Limited Data Set" per HIPAA definition. The clinical data and biospecimens are de-identified, and each subject is represented by a CBCP number. The link table is securely maintained at the office of the CBCP site PI which is strictly accessible only to the PI or his/her designee at that site. When clinical data Quality Assurance (QA) problems are identified in any participating institutions, they are all reported to WRAMC and the WRAMC clinical data team coordinates the QA resolving process. As of February 2011, over 5000 subjects have been enrolled in the study using two major questionnaires to collect up to 799 elements of clinicopathology data per subject. More than 43,000 specimens (including aliquots) have been collected and genomic and proteomic experiments have been conducted using these specimens.

GDP, on the other hand, is a consortium program constituting of nine clinical and research organizations with WRAMC serving as the lead institution. Each participating organization follows its own IRB-approved protocol allowing only "Safe Harbor Data Set", and each site manages and owns its own questionnaire-specific datasets. All the captured data are in a de-identified form, and each subject is represented by a GDP study ID. Each site PI maintains the link key to the identity of the subject enrolled at that site. All the de-identified data are then sent to WRI, and eventually loaded into the DW4TR. Aggregated subject information can be shared across the consortium members, but the detailed individual subject information in the DW4TR is only accessible to the originating clinical site. In GDP, about 500 subjects have been enrolled in the study. A dozen questionnaires were used to collect more than 7600 elements of data focusing on surgical procedures, pathology, family history, psychology, food intake, etc. Currently GDP is consolidating the questionnaires being used reducing the data elements to less than 1400.

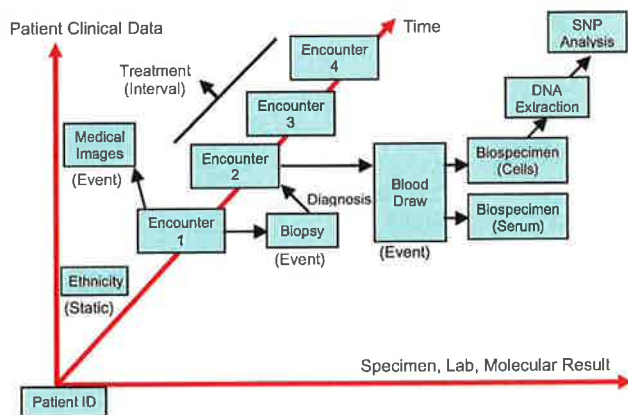


Fig. 1. Illustration in a 3-dimensional diagram of the relationship between clinical data (vertical axis), molecular data (horizontal axis), and temporal information (axis pointing into the page) represented by our data model. The patient is represented by Patient ID at origin. The three temporal data types are illustrated. This patient had clinical encounters at four time points, and "Encounter 1" resulted in medical images and biopsy which led to "Encounter 2", which resulted in biospecimens that could be used for different types of molecular studies. "Treatment" started from "Encounter 2" and ended at "Encounter 4".

3. Results

3.1. Overview of the DW4TR and the relationships between involved data types

Key to the development of the DW4TR is the understanding of the complex relationships between data collected or generated with multiple platforms. Fig. 1 illustrates our understanding of such relationships in a 3-dimensional patient data space which we can use to represent our data models.

Fig. 2 is a conceptual view of the DW4TR, as a hybrid system, that integrates the data collected or generated by the supported programs, and federates most of other data needed in the studies. Blocks with solid lines are the areas already developed or currently under development, and blocks with dashed lines are areas for future development. The whole DW is composed of a data tier, a middle tier, and an application tier. In the data tier, an extended EAV model is used for the raw clinical data. The middle tier is based on a patient-centric, modularly-structured clinical data model (including a medical image data model), integrating a specimen-centric molecular data model, and a temporal data model. The clinical and molecular data models are designed to be extensible. The application tier is composed of the ABB and the ISIV, and other utilities. Additional applications will be developed to federate and present other data needed but not generated by the supported translational research programs.

3.2. The ontology and the extensible data model

As a bottom-up approach, we developed a pragmatic light-weight ontology using controlled terminology both from existing ontologies and supplemental ones we developed to support the detailed program needs. The relationships between these terminologies in the ontology are reflected on the physical data models described below.

The patient-centric clinical data model was developed to reflect the physician-patient interaction. As illustrated in Fig. 3, the model is composed of six major modules: Medical History, Physical Exam, Diagnostic, Treatment, Outcome, and Scheduling and Consent. Each of these modules is composed of attributes and sub-modules; e.g., the Diagnostic module is composed of Biopsy, Imaging, and

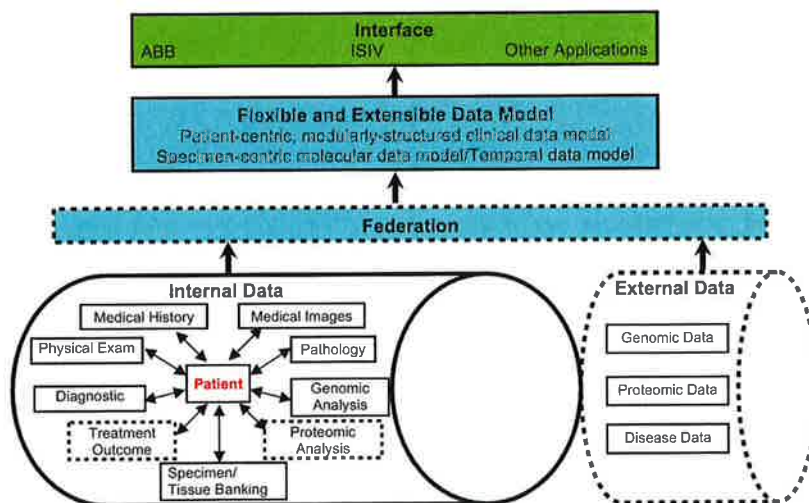


Fig. 2. The structure of the DW4TR including a data tier, a middle tier (blue), and an application tier (green). The data tier is composed of 'Internal Data' generated by the supported research project, and the 'External Data' not generated by but needed in the research. The "Internal Data" are integrated, and the "External Data" could be in an integrated or federated form but will be federated with the "Internal Data". Solid line blocks: developed or in development; dashed line blocks: future development. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

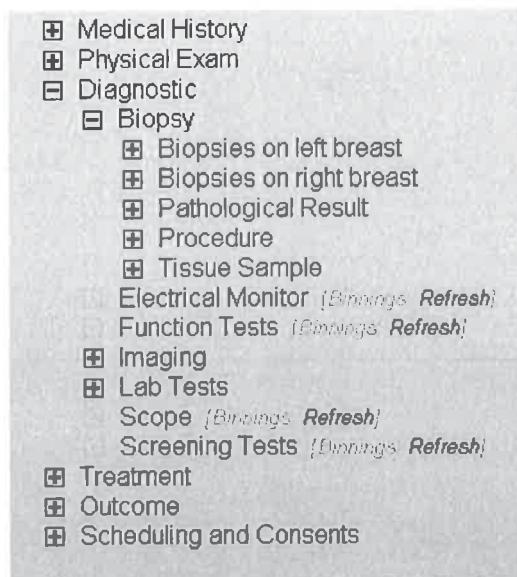


Fig. 3. A screenshot of the ontology structure of the patient-centric, modularly-structured clinical hierarchical data model.

Lab Tests, etc. Each of these sub-modules can then be further composed of its sub-modules and attributes. Many of these sub-modules are disease independent but some are disease specific. This modularly structured data model enables us to adapt the system to support the study of a new disease, by re-using disease independent modules, and developing a limited number of new modules specific to the new disease. The data model has a hierarchical structure. Users with domain knowledge can explore this hierarchical structure to retrieve needed data elements. Users with less specific domain knowledge may not be able to explore this data model as effectively though they can use the keyword search capability of the system to identify, for selection, the data elements containing keywords of interest.

3.2.1. The specimen-centric molecular study data model

The specimen-centric molecular study data model is connected to the clinical data model as sub-modules. For example, CBCP breast tissues and lymph nodes in different preservation types (e.g., OCT, FFPE, and FF) are represented by sub-modules of Diagnostic-Biopsy-Tissue Samples, blood samples and their derivatives (e.g., PAXgene tube, plasma, cells, serum, and clots) are represented by sub-modules of Diagnostic-Lab Tests. Genomic, proteomic, and other molecular studies are performed with molecular derivatives of the specimens (e.g. DNA, RNA, and protein) on different experimental platforms, e.g., IHC, FISH, gene expression microarray, SNP microarray, 2D-DIGE, MS, LC-MS, etc. We have completed the modules for the IHC and FISH assay data, and a proof-of-principle has been implemented for gene expression microarrays (c.f. Figs. 4 and 5).

3.2.2. Temporal information definition and presentation

All attributes are tagged with one of the three defined temporal properties. All temporal information, including static information, can be viewed in the ISIV (c.f. Fig. 7). Within the ABB users have the ability to define "time filters" for the different columns in the display (c.f. Fig. 6).

3.2.3. Example data module definition

Table 1 is a simplified illustration of three defined data modules. Thus, "Ethnic Group" is a simple (single attribute) data

module allowing multiple values for a given subject, and is "Static". "Alcohol Use" is a complex data module with three facets, and covers an "Interval". "Surgical Proc(EDURE)" is a module similar to "Alcohol Use" but is an "Event".

3.3. DW4TR interfaces—ABB and ISIV

3.3.1. ABB

This interface allows the user to analyze the data by dynamically creating a multidimensional pivot view. The rows are categorical data fields of interest, in a hierarchical structure. The columns can be data elements of either a numerical or categorical nature, in a flat or hierarchical structure. Although the view is a two dimensional table, the user can effectively explore data of theoretically unlimited dimensions by simply building multiple levels of a hierarchical data structure in both the rows and columns, and expanding them. The number of dimensions (levels of hierarchy) is only practically limited by the performance of the system (see Section 3.6). Fig. 4 (Panels A and B) shows screenshots of how an ABB view of five hierarchical levels was created and used to explore one subclass of breast cancer patients among CBCP Caucasian American subjects, who have triple-negative tumors, i.e., ER (estrogen receptor), PR (progesterone receptor), and HER2 (human epidermal growth factor receptor 2) negative.

Using the ABB, simple analyses such as mean, standard deviation, counts and percentages can be performed. Views on subject counts or specimen counts can be created. The results of interest can be printed, or exported to an Excel spreadsheet for additional analysis using other specialized software. A subject set of interest can also be saved as a cohort, for example the triple-negative Caucasian American breast cancer group shown in Fig. 4B, and used in a subsequent study or analyzed with a different application. The cohort can be analyzed using the ISIV described in the next section, or rendered to additional complicated analysis that the InforSense Analytical Workflow platform supports (not shown). The cohort data can also be exported to an Excel spreadsheet for analysis by specialized software outside of the DW4TR.

ABB offers many other features including a unique custom data binning capability. Multiple binning methods can be used, either automatically or manually, to create bins to explore data elements that contain either numerical or categorical values. For different studies, a user may need to use the same information in different ways. As shown in Fig. 4 (Panels C and D), the ethnic group attribute contains more than a dozen different possible values from CBCP and GDP. If the user wants to concentrate on Caucasians, African Americans, and Asian Americans, he/she can create three bins for them, and place the rest in "Other". A histogram of the binning can then be viewed. Custom data binning can be done with categorical data as well as numerical data, and in our practice viewing histograms helps to adjust the binning strategy to ensure that there are a sufficient number of records available in each bin.

One type of molecular data can be directly accessed through ABB is IHC. Fig. 4B shows data for three molecules (ER, PR and HER2) which are routinely assayed in the CBCP, for which there are established clinical standards for clinical IHC [76,77]. Two other molecules, Ki67 and p53, are also routinely assayed though there are no current clinical standards for these molecules. In CBCP, all five molecules are characterized by protein expression assayed by IHC, and HER2 is further characterized with gene copy number ratio assessed using FISH as needed. Both IHC and FISH data types are supported in the production system. The IHC submodule contains Result, Percentage, Intensity, and Staining Pattern for nuclear proteins (ER and PR) or Result and Score for membrane-bound proteins (HER2). The FISH submodule contains Result and Ratio. Both submodules belong to "Pathologic Results" of "Biopsy" of "Diagnostic" in the data model. Specific business rules are applied in

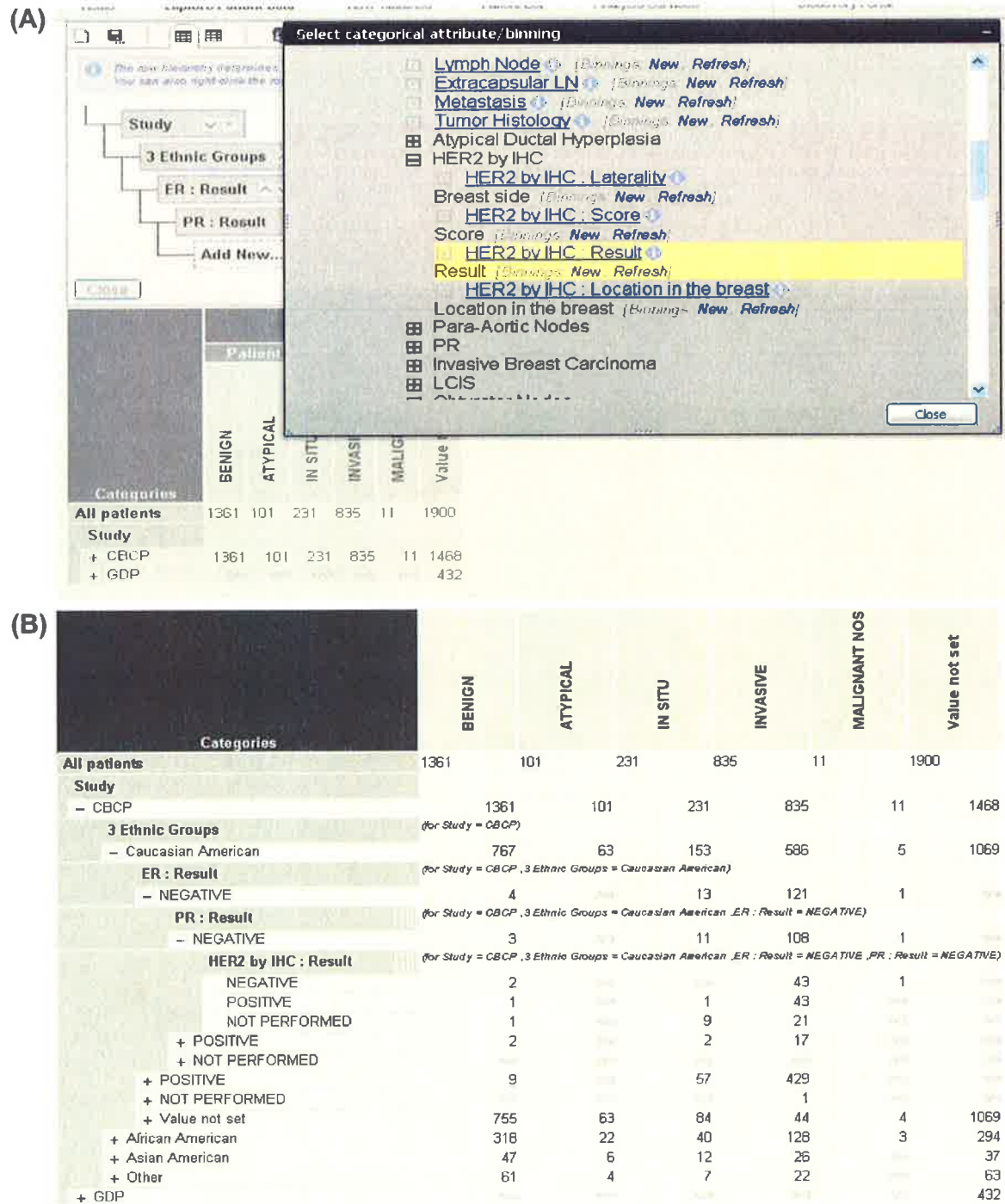


Fig. 4. Screenshots of ABB. (A) View construction. The column is 'Patient Pathology Category', and the rows are, hierarchically, 'Study', '3 Ethnic Group' (binned from 'Ethnic Group' as shown in Panels C and D), 'ER Result', 'PR Result', and the screen was captured when 'HER2 by IHC: Result' was being selected. The inset shows a portion of the data module where the data element resides. (B) View exploration. Rows of five levels were explored, and at the end the numbers are shown for triple negative (ER-, PR- and Her2-) subjects for CBCP Caucasian American in each pathology category. Similar analysis can be done for other ethnic groups. (C) The custom data binning feature. Here a manual binning option is selected. Bins created are 'Caucasian American', 'African American', 'Asian American', and 'Other', with Null defaulted. Available values are in the window of 'Unassigned Values', which can be dragged into the bins on the left to complete the binning. (D) Histogram of the binning result, which can be viewed during custom data binning or at the time of use.

the ETL process, for example the final HER2 expression result is determined by IHC Score (0 or 1+ as negative, and 3+ as positive), but when IHC Score is ambiguous (2+) the Ratio of the FISH assay is used, with ≤ 1.8 as negative and ≥ 2.2 as positive, and an intermediate ratio is considered as ambiguous.

A proof-of-principle model of microarray-based molecular data is in our test system; this feature is the only feature described in Section 3 that has not yet been applied to the production system. The MIAME standards [26] are followed with some modification, for example, our "sample ID" eliminates the need for subject and

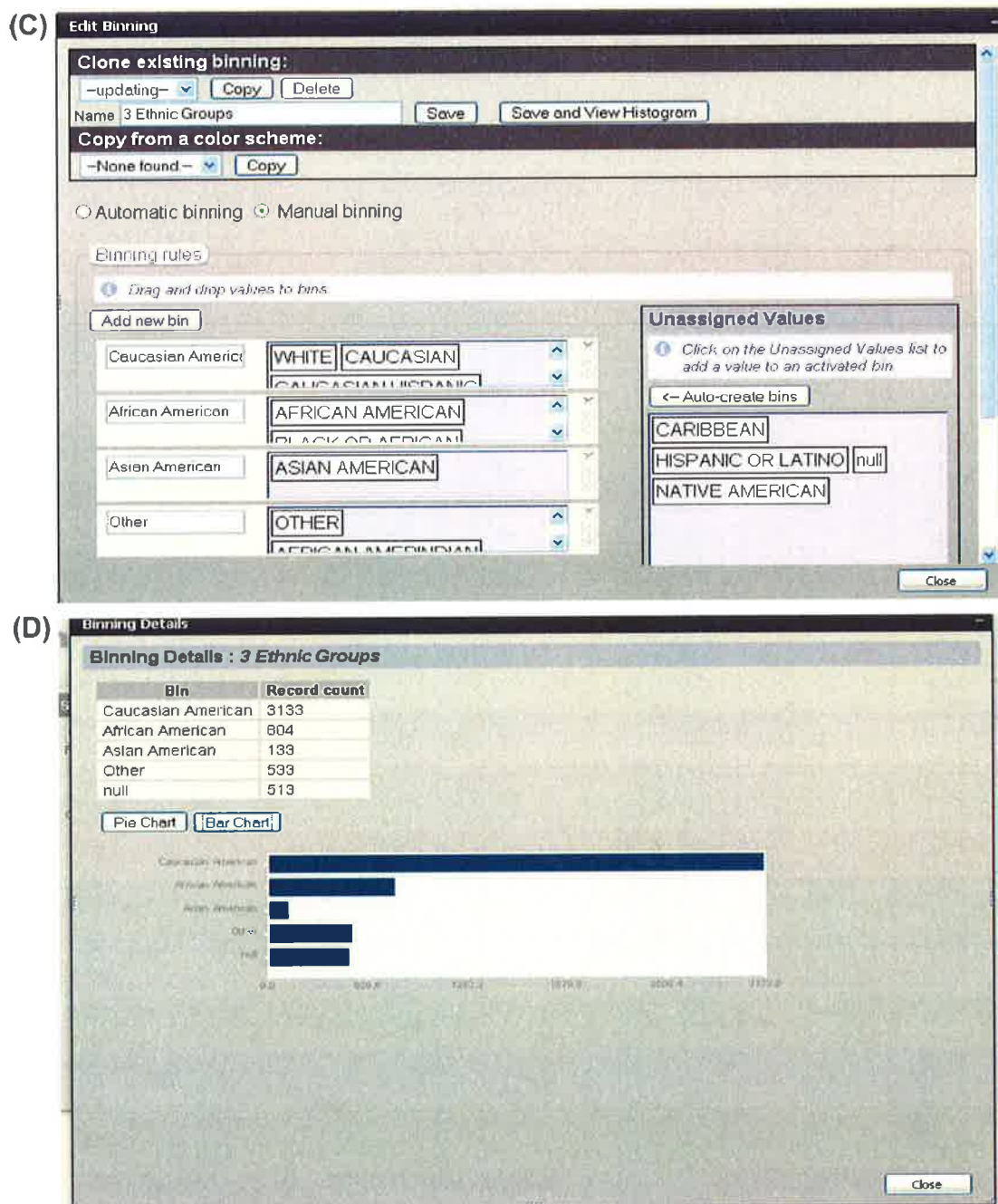


Fig. 4 (continued)

specimen information. The metadata are modeled, including sample ID, bio-molecule information (lab protocol, 260/280 ratio, RIN, etc.), array information (platform, array ID/batch number, etc.), experimental information (project name, PI, lab operator, date, etc.), and results (call rate, 3'/5' ratio, Pass/Fail flag, data file location, etc.). Metadata are accessible via ABB for selecting microarray data of interest in a cohort defined by clinicopathologic characteristics, but we saved the array raw data files in a file server due to their size and the fact that their analysis typically requires specialized software that read in the raw data file directly. Fig. 5 shows an example how a set of microarray data can be identified from previous experiments performed on samples from

a cohort of interest. This information may lead to a secondary use of the microarray data, or help the researcher to design a new experiment.

ABB also contains a utility "Time Filter" to apply temporal information in the analysis of aggregated information. The time filters are used to limit or constrain the data that are returned much like any other type of filter criteria. Time filters can be created based on absolute dates, patient age, or relative to other events. Fig. 6 shows one example how "Alcohol Usage" was analyzed for CBCP subjects when they were at different age ranges. "Alcohol Usage" is an Interval temporal data type with values self-reported by subjects covering different periods of time in patient's life. Note that direct

The screenshot shows the ABB software interface with a table titled 'Patient Count'. The table has columns for 'Patient Count', 'Patient Pathologic Category', and 'Define Time Filter'. The 'Patient Pathologic Category' column is expanded, showing sub-columns: ATYPICAL, BENIGN, IN SITU, INVASIVE, MALIGNANT NOS, and Value not set. The table lists various categories and their corresponding patient counts.

Categories	Patient Count	ATYPICAL	BENIGN	IN SITU	INVASIVE	MALIGNANT NOS	Value not set
All entities	4896	114	1607	301	1286	12	1620
2 Ethnic Groups							
- Caucasian American	3270	69	922	203	914	8	1183
(for 2 Ethnic Groups = Caucasian American)							
Menopausal Status							
- POST-MENOPAUSAL	1348	30	230	103	522	5	473
(for 2 Ethnic Groups = Caucasian American, Menopausal Status = POST-MENOPAUSAL)							
Microarray Experiment : Platform							
+ AFFY GENECHIP U133 PLUS2	138		3		76		61
Value not set	1210	30	227	103	446	5	412
- PRE-MENOPAUSAL	1283	24	484	66	235	3	480
(for 2 Ethnic Groups = Caucasian American, Menopausal Status = PRE-MENOPAUSAL)							
Microarray Experiment : Platform							
+ AFFY GENECHIP U133 PLUS2	46		2		19		26
Value not set	1237	24	482	66	216	3	454
+ STATUS POST HYSTERECTOMY	343	13	78	26	97		133
+ SURGICALLY MENOPAUSAL	379	11	86	25	126		140

Fig. 5. Screenshot of ABB showing exploration of available gene expression microarray data on biospecimens of a CBCP cohort of interest. Subjects were Caucasian American, post-menopausal or pre-menopausal, and experiments were performed on an Affymetrix platform. Subjects were mostly invasive breast cancer patients and normal subjects (as represented by 'Value Not Set' in the last column). The numbers from each pathology category do not completely add up to the total patient count on the left-most column since a couple of patients changed the category status with time. Highlighted row of subjects (light blue) can be saved as a cohort for further studies.

custom-binning of subject's age would not be useful to this analysis.

3.3.2. ISIV

This interface is for viewing and analysis of detailed subject information, most importantly the temporal-related information. It takes subject IDs as the input either by direct entry, or selecting from a pre-defined cohort. Events and Intervals of interest can be selected, and applied to the whole subject set. Information can be viewed independently for every subject, or aligned across the subject set on Events of interest to enable comparison between subjects. Static information can be displayed as well, so all the information about the subject can be studied. An example screenshot of ISIV is shown in Fig. 7.

Both the ABB and ISIV require minimal training before one can start to use them. The average training time for a clinician or scientist is 15 min, and the user can learn more features by using the system. Interestingly, for a high school student the average training time is only 10 min.

3.4. Direct data extraction from the database

The raw data in the DW4TR can be directly queried from the database, as a complete or partial dataset depending on the research need. The data can be presented in either a flat file structure or in Excel spreadsheet, for use with specialized data analysis software such as SAS, SPSS, or R. Such queries have been developed for and applied to both CBCP and GDP data. Direct data extraction requires a database administrator and is not the focus of this paper, thus is only briefly described here.

3.5. Extensibility of the DW4TR

The DW4TR was developed to support the CBCP with a vision to support translational research in general. This extensibility was tested and validated by expanding the system to support the GDP. Its flexibility was tested by our successful revision of the model commanded by the consolidation of GDP questionnaires and data elements. A number of disease-independent attributes and submodules developed for the CBCP implementation were re-used for the GDP instance, for example demographics and elements of past medical history. Many other attributes/submodules were modified to satisfy the requirements of GDP, for example alcohol usage. For GDP-specific data elements, additional modules were developed. In addition, the GDP contains a psychology questionnaire and a food questionnaire that are program-specific but not disease-specific. These were developed and structured to the overall patient-centric clinical data model, and could potentially be re-used for future programs. In addition, the GDP has different data security and accessibility requirements, and we have developed additional mechanism for it as detailed in Section 4. With the experience gained in supporting the GDP, we are confident that we will succeed in the current further extension of the system to support the third translational research program for the consortium led by Thomas Jefferson University.

3.6. System performance

To evaluate the performance potential of the DW4TR, we conducted a stress test using the CBCP data in a one-server system configuration. The test was done from a user's perspective, by

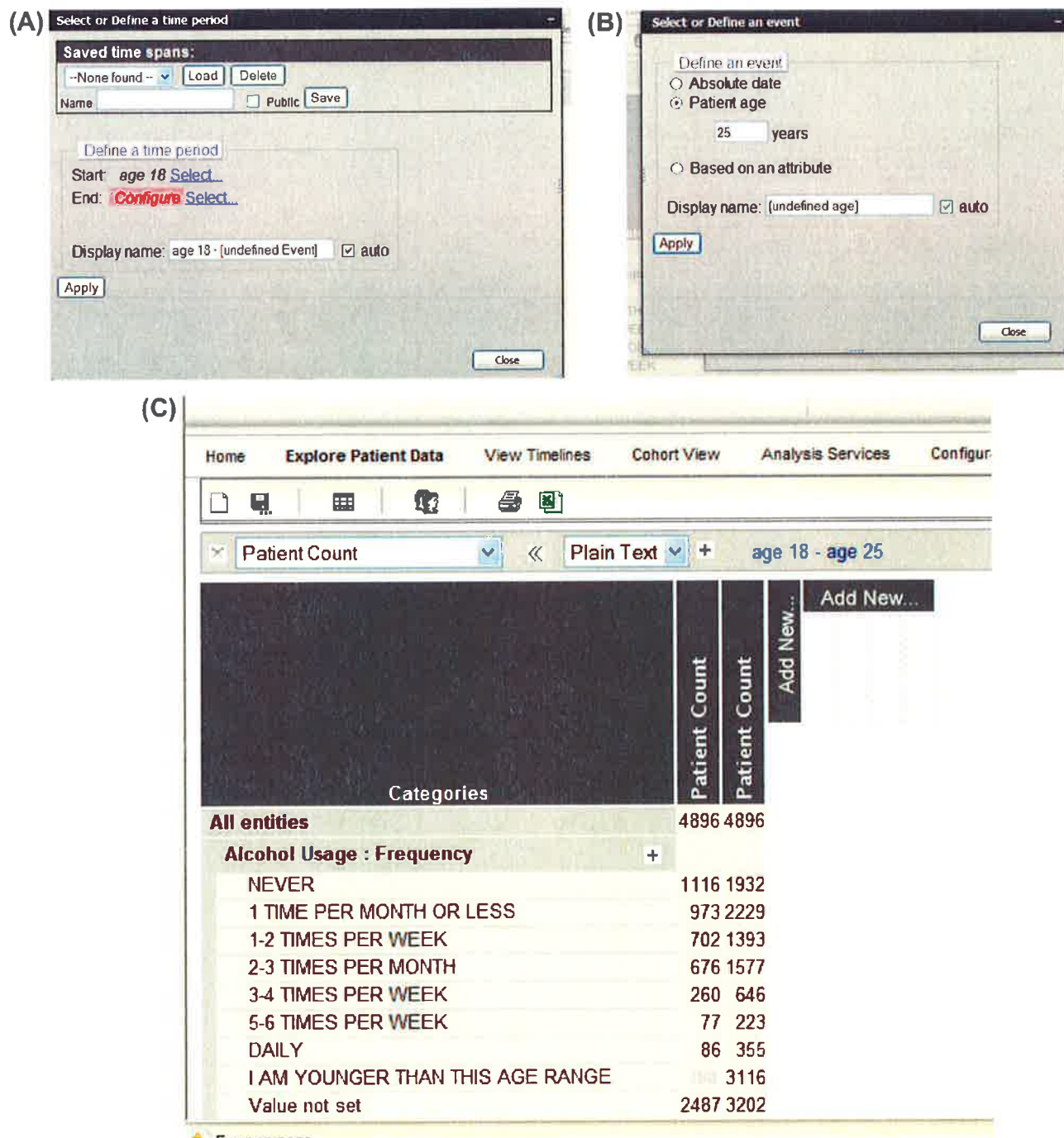


Fig. 6. Use of temporal information in ABB to study data of interests. (A) Screenshot showing the use of the "Time Filter" to define the starting and ending time for study, with starting time already being configured but not the ending time. (B) Configuring the ending time based on Age, but it can also be based on Absolute Date or an Attribute. (C) Example showing the alcohol usage history across the CBCP subjects. The first column shows the usage frequencies of subjects when they were between 18–25 years of age, and the second column shows the usage frequencies across the whole population. Additional age ranges can be configured to show the corresponding alcohol usage habit.

measuring how long it takes for the data to completely display in a hierarchical view for the first time—note that subsequent display of the same data is much faster. The view is composed of three levels of Pathology Category, Ethnicity, and Body-Mass Index (BMI, with custom-binning) in the rows and record counts in the columns. Starting from the then patient number of 4234, we artificially replicated the records in the database to 8468, 16,936, 33,872, and 67,744 without performing a database tune-up other than Oracle table analysis. It took one second or less, for the level

1 (Pathology Category) and level 2 (Ethnicity) data to be displayed in all these tests. For level 3 (BMI), it took 2, 4, 12, 39, and 134 s respectively to display the data when the number of records doubled stepwise from 4234 to 67,744. Given that the current CBCP annual subject enrollment is about 600, it will take another 20 years before 16,936 subjects are enrolled when 12 s are needed to display 3 levels of data. We expect that the system performance will be further dramatically improved with hardware upgrades and Oracle tuning.

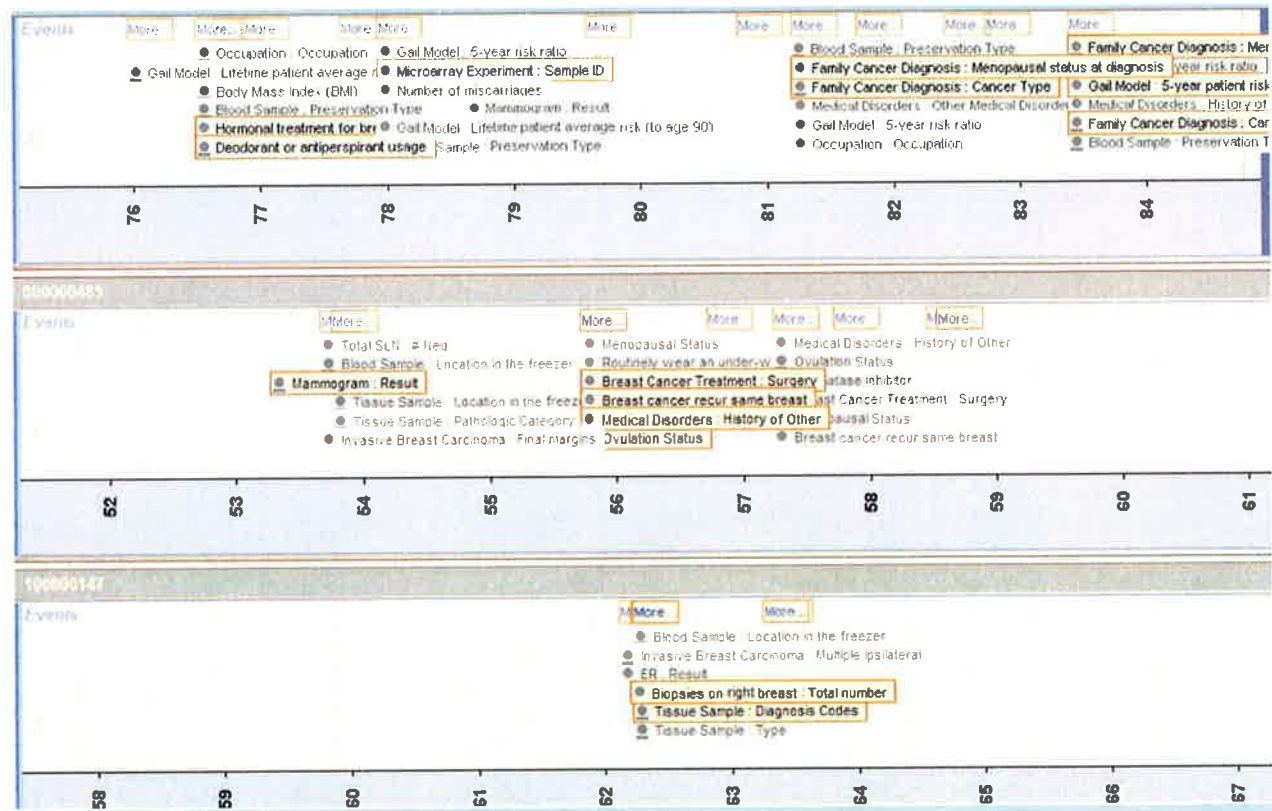


Fig. 7. Screenshot of the ISIV. Temporal information from three subjects is shown. Users can zoom in or out, align the information on a specific event for cross-subject comparison, and explore more detailed information by clicking on the event of interest. The temporal axis shows the age but can be toggled to calendar year.

Table 1
Example data module definition.

ID	Attribute	Facet	Data type	Units	Multiple values	From date	To date	Temporal information	Parent ID
23	Ethnic Group		VARCHAR2		Yes			Static	
103	Alcohol Use		COMPLEX		No	mm/dd/yyyy	mm/dd/yyyy	Interval	
104	Alcohol Use	Name	VARCHAR2		No	mm/dd/yyyy	mm/dd/yyyy	Interval	103
105	Alcohol Use	Frequency	NUMBER	/day	No	mm/dd/yyyy	mm/dd/yyyy	Interval	103
106	Alcohol Use	Amount	NUMBER	drinks	No	mm/dd/yyyy	mm/dd/yyyy	Interval	103
1541	Surgical Proc		COMPLEX		No	mm/dd/yyyy		Event	
1542	Surgical Proc	Name	VARCHAR2		No	mm/dd/yyyy		Event	1541
1543	Surgical Proc	Code	VARCHAR2		No	mm/dd/yyyy		Event	1541
1544	Surgical Proc	Laterality	VARCHAR2		No	mm/dd/yyyy		Event	1541

3.7. Example use cases

The DW4TR has been playing an important role in supporting translational research across user groups in both WRI and WRAMC. Here we report four example use cases to demonstrate the utility of the system.

3.7.1. The DW4TR enables clinicians to study the integrated data directly

When first exposed to the DW4TR, CBCP clinical users appreciated the simple interface which gave them their first opportunity to directly study the integrated clinicopathologic and biospecimen data. They were happy to see that the results shown on the ABB confirmed their qualitative clinical observations. In one short session of less than 1 h, the program PI identified several lines of evidence which subsequently led to an abstract accepted to the San Antonio Breast Cancer Symposiums [80]. The system has since

been further improved in both the data model structure and the user interface.

3.7.2. The DW4TR enables enhanced cohort and biospecimen selection

Before the DW4TR was developed, cohort and biospecimen selection for any research project was a major task involving multiple data sources including: Core Questionnaires, Pathology Checklists, and IHC assay reports, in the form of an Oracle database, a Microsoft Access database, Excel spreadsheets, and occasionally hard copy reports. Many manual steps were involved, and the procedure was not standardized. The selection of biospecimens was especially challenging when there were temporal restrictions, for example identifying blood samples drawn in a subject cohort before certain kinds of invasive procedures were performed. This requirement alone could take an experienced researcher several days since multiple questionnaires are involved and a number of surgical procedures could be performed on one subject and that there were 15

Table 2

Example classifications of breast neoplasms in MeSH, SNOMED, and DW4TR.

Name	Classifications
MeSH1	3. Diseases (C) → Skin and Connective Tissue Diseases [C17] → Skin Diseases [C17.800] → Breast Diseases [C17.800.090] → Breast Neoplasms [C17.800.090.500]
MeSH2	3. Diseases (C) → Neoplasms [C04] → Neoplasms by Site [C04.588] → Breast Neoplasms [C04.588.180]
SNOMED1	SNOMED CT Concept (SNOMED RT + CTV3) → Clinical finding (finding) → Disease (disorder) → Disorder by body site (disorder) → Disorder of body system (disorder) → Disorder of breast (disorder) → Neoplasm of breast (disorder)
SNOMED2	SNOMED CT Concept (SNOMED RT + CTV3) → Clinical finding (finding) → Finding by site (finding) → Finding of body region (finding) → Finding of trunk structure (finding) → Finding of region of thorax (finding) → Breast finding (finding) → Disorder of breast (disorder) → Neoplasm of breast (disorder)
DW4TR	Medical History → Past Medical History → Major Adult Illnesses → Breast Cancer

different kinds of such procedures. Using the DW4TR interface and additional utilities, the time needed to perform such tasks has now been reduced to minutes. Currently any cohort and biospecimen selection can be done in hours instead of days or weeks.

3.7.3. The DW4TR serves as a research environment for risk factor assessment

The DW4TR provides an effective research environment for cancer risk factor analysis. Several studies examining breast cancer risk factors have been done using this system [81–85], and several new studies are underway.

3.7.4. The DW4TR facilitates virtual experimental studies

Virtual experimental studies can be performed as integrated information associated with completed research projects is available. Using the information stored in the DW4TR and the file server, we identified the gene expression microarray data from three previous projects, on two types of specimens from three subject groups. From these datasets we were able to create two virtual experiments, and both studies were accepted to leading scientific conferences [86,87].

4. Security and accessibility

Data security and access control are important factors in the design of the DW4TR, particularly as our system can be expanded to include multiple research studies. The use of clinical data in translational research is governed by HIPAA compliant IRB approved protocols, and typically cannot be unconditionally shared. Thus we have designed and developed a set of data security and access solutions to satisfy such requirements.

The first layer of security for the system is data de-identification. Prior to being loaded into the DW4TR, PHI is stripped away from the collected data. At this point subjects will have been assigned a unique research subject ID for linking together subject information across multiple sources. The next level of security makes use of the “Virtual Private Database” (VPD) feature of Oracle, which allows for row level access control within the DW. While one of the strengths of the system is the ability to perform cross-study analysis, in many instances it is necessary to restrict data visibility among different users of the system. In our security model, all users are assigned roles which dictate the data they are permitted to access. The VPD system is designed such that every data element in the DW is tagged with a study identifier that is associated with the different user roles, enabling data access control. The third level of security is at the application tier. All access to the DW occurs through a password protected web portal that is served by an application server. The final layer of security is a network firewall and Virtual Private Network (VPN) for which the 168-bit encryption Secure Sockets Layer VPN from Juniper Networks Inc. is deployed. All components of the system, including the DW4TR and the web portal, are protected behind the local network firewall and are only accessible from either within the Institute itself, or by first establishing a VPN connection to the network.

These security and access control measures were developed based on the needs of the supported programs, and can be selectively applied to individual programs. For the two programs currently supported by the DW4TR, CBCP protocols allow data access by all CBCP researchers, and VPN access to the DW4TR is granted with written permission of the CBCP program PI. For GDP, aggregated biomedical information is allowed for use by all the GDP researchers, but the detailed subject information is only allowed to the originating clinical sites unless additional paperwork is completed. Thus, we have designed and developed a role-based site-specific privilege system to enable GDP users to use ABB and the ISIV. DW4TR can also be configured to disable functions and utilities of choice. Currently at the request of the GDP PI, we have disabled the ISIV functions for GDP users so that the GDP data are only available through ABB to GDP users for aggregated information. When a GDP researcher identifies a cohort of interest, additional data use agreements will be completed to request individual subjects' information from the corresponding clinical sites, for approval by the involved sites and the GDP program PI. With a written request from the GDP program PI, WRI will supply the detailed data from the DW4TR to the researcher.

5. Discussion

We have taken a bottom-up approach to develop a DW4TR to support translational research programs. The system is extensible, integrating internally generated clinicopathologic and molecular data based on a light-weight ontology, with a patient-centric, modularly structured clinical data model supplemented with a specimen-centric molecular data model. Two interfaces have been developed targeting non-informatician end users, with the ABB supporting aggregated data analysis and ISIV supporting single subject information analysis. The system enables study cohort and biospecimen selection and is capable of handling temporal relationships between clinical data and biospecimen data collected at multiple points in time.

The development of our data model relies on using or creating ontologies for clinical and molecular data types. We will focus this discussion on clinical data ontology. In fact, for molecular data types, the scale of work is much smaller and the existing ontologies are mostly adequate. For example, we found that the MIAME standard met our needs for modeling gene expression microarray metadata. Our analysis indicated that existing ontologies for clinical data cannot be readily adapted to satisfy our specific research needs, due to gaps between existing ontologies and our needs as well as overlaps and inconsistencies between existing standards and ontologies [70,71]. For example, the CBCP Pathology Checklist contains 372 data elements including 131 breast pathology conditions. When the data form was developed and revised, the AJCC (American Joint Committee on Cancer) guidelines and ASCO/CAP (American Society of Clinical Oncology/College of American Pathologists) guidelines were followed; these guidelines, like many other guidelines, have evolved over the years [88–91]. Many data elements present in the CBCP lack standard descriptions, for example

for many benign diseases. Some biomarkers used in the CBCP (Ki67 and p53) do not even have a standard clinical IHC protocol. Therefore, descriptions or definitions of such data elements were developed based on medical textbooks and pathologist's clinical practice for the Pathology Checklist, which we modeled ourselves.

There are classifications in existing standards/ontologies that we find cumbersome or which do not meet our needs [92]. We show as an example in Table 2 the classifications of "Breast Neoplasms" by two existing ontologies. In MeSH, it is classified under "Skin Diseases". This classification (named "MeSH1" in Table 2) is not available in the SNOMED-CT classifications under either Skin Diseases or Connective Tissue Diseases. In MeSH, "Breast Neoplasms" is also classified under "Neoplasms" (in a way we agree with, "MeSH2" in Table 2). However, in the two classifications, "Breast Neoplasms" have different codes, which may cause a problem in the practical applications of data modeling. In SNOMED-CT, the term "Neoplasm of Breast" is used, and there are several classifications leading to it with relatively long paths. We show two of them, "SNOMED1" and "SNOMED2" in Table 2. In our pragmatic approach to support the CBCP, applying such a complete classification is cumbersome. In addition, reconciliation of discrepancies among the available ontologies and our understanding of the problem, for all the data elements in the CBCP, is a major task and not one of our immediate project needs. Therefore, we have developed a classification path for "Breast Cancer", which is more light-weight but adequate to our needs (see "DW4TR" in Table 2).

Finally, some questions in the questionnaires of the two programs we currently support do not comply with existing standards, but we are still required to model them to enable proper storage and presentation of the data. Such problems have been reported to program PIs and were taken into consideration, for example in the consolidation of the GDP questionnaires. In addition, existing ontologies are not static, for example MeSH is currently updated weekly and SNOMED-CT is updated monthly. Thus, after analyzing what is available and what we need to support, we determined that our best approach was to design a light-weight ontology incorporating existing standards and ontologies where possible and to map our classification system to existing ones at a later time. The mapping table(s) will serve as the buffer between the two moving targets which will ensure transportability and un-interrupted use of the system. We will also supplement those standards with new ontologies and common data elements that we define.

The DW4TR currently supports limited number of molecular data types and we will continue to develop data models to support other molecular study platforms, including SNP microarray and array CGH data. We expect that the method developed for gene expression microarray will in general be applicable to these array-based technologies. We will explore support for RT-PCR, tissue microarrays, mass spectrometry, and Next Generation sequencing technologies. These will depend on the needs of the users and programs that we support.

Currently a prototype has been developed to enable access to medical and molecular images through the ISIV, including digital/digitized mammograms and gene expression microarray images. A thumbnail of the image is stored in the database of the DW with searchable annotations. The full-size image is stored in a file server instead of the DW4TR database for performance considerations. We plan to follow the DICOM standard for medical images and will explore how to best annotate molecular images. Additional images we need to support include, pathology H&E slide images, biomarker IHC images, and tissue microarray images. Note that all the medical images have to be de-identified for research purpose.

Currently the DW4TR does not support EMR for data input. We have a stepwise plan to begin using EMR records for CBCP subjects and we are looking at subject de-identification, data extraction, and data model expansion. Although we do not plan to take a

systematic "Enterprise" approach, we will learn from such approaches as used by I2B2 and others. When EMRs are used as an input, a number of functions need to be integrated into the system including subject de-identification and information extraction, for which the I2B2 development teams have made good progresses [50,54,93]. While I2B2 clearly plays an important role in the study design phase including addressing de-identification and natural language processing, it was not designed to be a comprehensive solution that can provide the level of detailed access necessary for later stage research [50–54]. The ABB described here provides the same ability to generate aggregate statistics for population stratification and cohort selection, but by doing this in a web-based pivot table interface it also facilitates trend identification and discovery in a way traditional query interfaces cannot. From a data modeling perspective, the ABB also allows for more sophisticated handling of complex, multi-faceted attributes whereas I2B2 only offers aggregation around the single pivot point of the patient. I2B2 makes good use of ontologies, but has not yet developed an ontology with the level of detail described here [54,94]. Furthermore, by design I2B2 does not enable access to row level clinical data and detailed clinical timeline provided by the ISIV [54,94].

In conclusion, we have developed the DW4TR to support a breast cancer translational research program and then expanded it to support a gynecological disease translational research program. It is based on a light-weight ontology and equipped with an interface capable of handling temporal information, designed for use by non-informatician specialists including clinicians and laboratory scientists. The system supports both *in silico* and *in vitro* studies. With its proven extensibility, we believe that the DW4TR will play an important role in translational research across multiple disease studies.

Acknowledgments

We thank Mr. Anton Oleynikov, Mr. Faruk Cay, Mr. Raveen Sharma, Dr. Csaba Mihaly, Mr. Eric Babyak, and Dr. Yonghong Zhang, for their help in this project. We thank Ms. Joni Kohr, Dr. Hallgeir Rui, and Mr. John Eberhardt for their help in manuscript preparation and revision. We also thank the anonymous reviewers for their constructive comments that resulted in enhanced quality of this paper. This work was supported by the Clinical Breast Care Project, and the Gynecological Disease Program, with funds from the US Department of Defense through Henry Jackson Foundation for the Advancement of Military Medicine, Rockville, MD. The views expressed in this paper are those of the authors and do not reflect the official policy of the Department of the Army, Department of Defense, or the government of the United States.

References

- [1] Hu H, Mural RJ, Liebman MN, Hu H, Mural RJ, Liebman MN, et al. Biomedical informatics in translational research. Boston, London: Artech House; 2008.
- [2] Bernstam EV, Hersh WR, Johnson SB, Chute CG, Nguyen H, Sim I, et al. Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical and Translational Science Award programs. *Acad Med* 2009;84:964–70.
- [3] Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. *Physiol Genomics* 2009;39:131–40.
- [4] Sarkar JN. Biomedical informatics and translational medicine. *J Transl Med* 2010;8:22.
- [5] Nadkarni PM, Reeders ST, Zhou J. CECIL: a database for storing and retrieving clinical and molecular information on patients with Alport syndrome. In: *Proc annu symp comput appl med care*; 1993. p. 649–53.
- [6] Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, Xiong R, et al. Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. *Pharmacogenomics* 2004;5:933–41.
- [7] Brammen D, Katzer C, Rohrig R, Weismuller K, Maier M, Hossain H, et al. An integrated data-warehouse-concept for clinical and biological information. *Stud Health Technol Inform* 2005;116:9–14.

- [8] Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. In: AMIA annu symp proc; 2006. p. 1040.
- [9] Hu H, Correll M, Osmond M, Gao J, Oleynikov A, Sheldon J, et al. A clinical data warehouse to support translational research. In: 15th annual international conference on Intelligent Systems for Molecular Biology (ISMB), Vienna, Austria; July 21–25, 2007.
- [10] Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform 2008;91:101.
- [11] Rossille D, Burgun A, Pangault-Lorho C, Fest T. Integrating clinical, gene expression, protein expression and preanalytical data for in silico cancer research. Stud Health Technol Inform 2008;136:455–60.
- [12] Hu H. Data centralization. In: Hu H, Mural RJ, Liebman MN, editors. Biomedical informatics in translational research; 2008. p. 141–71 [chapter 8].
- [13] Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – an integrated standards-based translational research informatics platform. In: AMIA annu symp proc; 2009. p. 391–5.
- [14] Inmon WH, Kelley C. Developing the data warehouse. QED Publishing Group; 1993.
- [15] Kimball R, Ross M. The data warehouse Toolkit. 2nd ed. John Wiley and Sons; 2002.
- [16] Stead WW, Hammond WE, Straube MJ. A chartless record – is it adequate? J Med Syst 1983;7:103–9.
- [17] Niedner CD. The entity-attribute-value data model in radiology informatics. In: Proceedings of the 10th conference on computer applications in radiology, Anaheim, CA; 1990.
- [18] Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc 1998;5:511–27.
- [19] Codd F, Codd SB, Salley CT. Providing OLAP (Online Analytical Processing) to user-analysts: an IT mandate. San Jose: Codd & Date, Inc; 1993.
- [20] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2008;9:75–90.
- [21] Health Level 7. <<http://www.hl7.org/>>.
- [22] Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 1994;271:1103–8.
- [23] SNOMED Clinical Terms® (SNOMED CT®). <http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html>.
- [24] DICOM – Digital Imaging Communications in Medicine. <<http://www.medical.nema.org/>>.
- [25] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25–9.
- [26] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.
- [27] BioPAX: Biological Pathways Exchange. <<http://www.biopax.org/>>.
- [28] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40:30–43.
- [29] Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS 2006;10:185–98.
- [30] Cote RG, Jones P, Apweiler R, Hermjakob H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinform 2006;7:97.
- [31] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993;32:281–91.
- [32] Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. Nat Rev Genet 2004;5:213–22.
- [33] Yu AC. Methods in biomedical ontology. J Biomed Inform 2006;39:252–66.
- [34] Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform 2006;124:35.
- [35] Weedon-Fekjaer H, Lindqvist BH, Vatten LJ, Aalen OO, Tretli S. Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res 2008;10:R41.
- [36] Shoham Y. Reasoning about change: time and causation from the standpoint of artificial intelligence. Cambridge, MA, USA: MIT Press; 1988.
- [37] Shahar Y. A framework for knowledge-based temporal abstraction. Art Intell 1997;90:79–133.
- [38] Deshpande AM, Brandt C, Nadkarni PM. Temporal query of attribute-value patient data: utilizing the constraints of clinical studies. Int J Med Inform 2003;70:59–77.
- [39] Ceusters W, Steurs F, Zanstra P, Van Der Haring E, Rogers J. From a time standard for medical informatics to a controlled language for health. Int J Med Inform 1998;48:85–101.
- [40] Das AK, Shahar Y, Tu SW, Musen MA. A temporal-abstraction mediator for protocol-based decision-support systems. In: Proc annu symp comput appl med care; 1994. p. 320–4.
- [41] Shahar Y, Das AK, Tu SW, Kraemer FB, Musen MA. Knowledge-based temporal abstraction for diabetic monitoring. In: Proc annu symp comput appl med care; 1994. p. 697–701.
- [42] Nguyen JH, Shahar Y, Tu SW, Das AK, Musen MA. A temporal database mediator for protocol-based decision support. In: Proc AMIA annu fall symp; 1997. p. 298–302.
- [43] Post AR, Harrison Jr JH. PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. J Am Med Inform Assoc 2007;14:674–83.
- [44] Post AR, Sovarel AN, Harrison JH, Jr. Abstraction-based temporal data retrieval for a Clinical Data Repository. In: AMIA annu symp proc; 2007. p. 603–7.
- [45] Chen J, Zhao P, Massaro D, Clerch LB, Almon RR, DuBois DC, et al. The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface. Nucleic Acids Res 2004;32:D578–81.
- [46] Yamamoto Y, Namikawa H, Inamura K. Development of a time-oriented data warehouse based on a medical information event model. Igaku Butsuri 2002;22:327–33.
- [47] Informatics for Integrating Biology and the Bedside (I2B2). <<https://www.i2b2.org/>>.
- [48] Mendis M, Wattanasin N, Kuttan R, Pan W, Phillips L, Hackett K, et al. Integration of hive and cell software in the i2b2 architecture. In: AMIA annu symp proc; 2007. p. 1048.
- [49] Mendis M, Phillips LC, Kuttan R, Pan W, Gainer V, Kohane I, et al. Integrating outside modules into the i2b2 architecture. In: AMIA annu symp proc; 2008. p. 1054.
- [50] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14:550–63.
- [51] Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records – an experience report. Stud Health Technol Inform 2009;143:441–6.
- [52] Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol 2009;9:70.
- [53] Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. In: AMIA annu symp proc; 2009. p. 442–6.
- [54] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124–30.
- [55] The stanford translational research integrated database environment (STRIDE). <<http://clinicalinformatics.stanford.edu/STRIDE/>>.
- [56] Hernandez P, Podchyska T, Weber S, Ferris T, Lowe H. Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse. AMIA Annu Symp Proc 2009;2009:244–8.
- [57] Lowe HJ, Huang Y, Regula DP. Using a statistical natural language parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. In: AMIA annu symp proc; 2009. p. 386–90.
- [58] calIntegrator. <<https://cabig.nci.nih.gov/tools/calIntegrator>>.
- [59] Repository of Molecular Brain Neoplasia Data (REMBRANDT). <<http://caintegrator-info.nci.nih.gov/rembrandt/>>.
- [60] Cancer Genetic Markers of Susceptibility (CGEMS). <<http://cgems.cancer.gov/>>.
- [61] Clinical and Translational Science Award (CTSA), on the website of National Center for Research Resources. <http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards>.
- [62] The Cancer Biomedical Informatics Grid (caBIG). <<http://cabig.nci.nih.gov/>>.
- [63] von Eschenbach AC, Buetow KH. Cancer informatics vision: caBIGTM. Cancer Inform 2006;2:22–4.
- [64] The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Stud Health Technol Inform 2007;129:330–4.
- [65] Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragos G, Sahni H, et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19:2404–12.
- [66] Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragos G, et al. caCORE version 3: implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2008;41:106–23.
- [67] McCusker JP, Phillips JA, Gonzalez Beltran A, Finkelstein A, Krauthammer M. Semantic web data warehousing for caGrid. BMC Bioinform 2009;10(Suppl. 10):S2.
- [68] The Clinical Breast Care Project. <www.cbcp.info/>.
- [69] The Gynecological Disease Program. <<http://www.gyndisease.org/index.html>>.
- [70] Rector AL, Zanstra PE, Solomon WD, Rogers JE, Baud R, Ceusters W, et al. Reconciling users' needs and formal requirements: issues in developing a reusable ontology for medicine. IEEE Trans Inf Technol Biomed 1998;2:229–42.
- [71] Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc 2007;14:687–96.
- [72] Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, Schmandt L, et al. The development and deployment of Common Data Elements for tissue banks for translational research in cancer – an emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC Cancer 2008;8:91.
- [73] Min H, Manion FJ, Goralczyk E, Wong YN, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. J Biomed Inform 2009;42:1035–45.
- [74] caBIG Vocabularies & Common Data Elements (VCDE) Workspace. <<https://cabig.nci.nih.gov/workspaces/VCDE/>>.
- [75] Brandt C, Cohen DB, Shifman MA, Miller PL, Nadkarni PM, Frawley SJ. Approaches and informatics tools to assist in the integration of similar clinical research questionnaires. Methods Inf Med 2004;43:156–62.
- [76] Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, et al. American Society of Clinical Oncology/College of American Pathologists

- guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med* 2007;131:18–43.
- [77] Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Arch Pathol Lab Med* 2010;134:907–22.
- [78] Beaulah SA, Correll MA, Munro RE, Sheldon JG. Addressing informatics challenges in Translational Research with workflow technology. *Drug Discov Today* 2008;13:771–7.
- [79] Hu H, Kvecher L. Data tracking systems. In: Hu H, Mural RJ, Liebman MN, editors. *Biomedical informatics in translational research*; 2008. p. 111–36 [chapter 7].
- [80] Hu H, Zhang Y, Kvecher L, Sun W, Hooke J, Mural RJ, et al. Different characteristics of invasive breast cancers between Caucasian and African American women. In: *The 29th San Antonio breast cancer symposium*, San Antonio, TX; 14–17 December 2006.
- [81] Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN. A Bayesian derived network of breast pathology co-occurrence. *J Biomed Inform* 2008;41:242–50.
- [82] Bekhash A, Maskery SM, Kvecher L, Correll M, Zhang Y, Hooke J, et al. Clinical breast care project data warehouse as a research environment for breast cancer risk factor studies; submitted for publication.
- [83] Bekhash A, Maskery SM, Kvecher L, Hooke J, Liebman MN, Shriver CD, et al. A pilot study of known or controversial breast cancer risk factors using the Clinical Breast Care Project database as a research environment. In: *The 30th San Antonio breast cancer symposium*, San Antonio, TX; 13–16 December 2007.
- [84] Saini J, Kovatich A, Bekhash A, Hooke J, Mural RJ, Shriver CD, et al. Association of clinicopathologic characteristics with IHC-based breast cancer subtypes. *Cancer Res* 2009;69:635s.
- [85] Bekhash A, Saini J, Li X, Rapuri P, Hooke AJ, Mural RJ, et al. Ethnicity difference of benign breast diseases in breast cancer and non-cancer patients. *Cancer Res* 2010;288s–9s.
- [86] Saini J, Li X, Kvecher L, Larson C, Croft D, Yang YC, et al. Differential gene expression analysis among post-menopausal caucasian invasive breast cancer, benign and normal subjects. *Cancer Res* 2010;245s.
- [87] Li X, Rapuri P, Melley J, Brilhart G, Wu W, Kvecher L, et al. Comparative analysis of gene expression profiles in human breast cancer from microarray data using breast tissues and peripheral blood samples. In: *International conference on intelligent systems for molecular biology (ISMB)*, Boston, MA; 11–13 July 2010.
- [88] Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A. *AJCC cancer staging manual*. 7th ed. New York, NY: Springer-Verlag; 2010 [6th printing].
- [89] Lester SC, Bose S, Chen YY, Connolly JL, de Baca ME, Fitzgibbons PL, et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch Pathol Lab Med* 2009;133:1515–38.
- [90] Lester SC, Bose S, Chen YY, Connolly JL, de Baca ME, Fitzgibbons PL, et al. Protocol for the examination of specimens from patients with ductal carcinoma in situ of the breast. *Arch Pathol Lab Med* 2009;133:15–25.
- [91] Printz C. New AJCC cancer staging manual reflects changes in cancer knowledge. *Cancer* 2010;116:2–3.
- [92] Andrews JE, Richesson RL, Krischer J. Variation of SNOMED CT coding of clinical research concepts among coding experts. *J Am Med Inform Assoc* 2007;14:497–506.
- [93] Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *J Am Med Inform Assoc* 2009;16:571–5.
- [94] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. In: *AMIA annu symp proc*; 2007. p. 548–52.